The Dissertation Committee for Christopher Dale White
certifies that this is the approved version of the following dissertation:

Optimality Guarantees for Non-convex Low Rank Matrix Recovery Problems

Committee:

Rachel Ward, Supervisor

Francois Baccelli

Braxton Osting

Sujay Sanghavi

Mihai Sirbu

Gordan Zitkovic

Optimality Guarantees for Non-convex Low Rank Matrix Recovery Problems

by

Christopher Dale White, B.S., B.A.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2015

Dedicated to Aunt Anne, Uncle Dale and Grandmommy.

Acknowledgments

I would first like to acknowledge my friends and family for their support over the past 6 years and in particular, the good friends that I made during grad school. A special thanks to my fellow plaid-wearing colleagues Andrew Kontaxis, Matthew Pancia, Aaron Royer and Sam Taylor for all of their helpful conversations. I am also forever grateful to the support and tolerance of my parents, Shannon Hatfield and Adele Powers.

I would also like to thank my dissertation committee for their valuable input and time. A huge special thanks to my advisor Rachel Ward for putting up with me for so many years; I am extremely grateful for her patience and her continued assistance. I would also like to thank Andrea Bertozzi for generously supporting my travel and work at UCLA. A special thanks also to Braxton Osting for his guidance and mentorship during my time at UCLA and after.

Lastly I would like to thank the UT Mathematics department for their financial support and resources throughout my time at UT. In particular, I am indebted to Dan Knopf, Thomas Chen, Elisa Bass and Eva Hernandez for their help.

Optimality Guarantees for Non-convex Low Rank Matrix Recovery Problems

Publication No. _____

Christopher Dale White, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Rachel Ward

Low rank matrices lie at the heart of many techniques in scientific computing and machine learning. In this thesis, we examine various scenarios in which we seek to recover an underlying low rank matrix from compressed or noisy measurements. Specifically, we consider the recovery of a rank r positive semidefinite matrix $XX^T \in \mathbb{R}^{n\times n}$ from m scalar measurements of the form $y_i := a_i^T XX^T a_i$ via minimization of the natural ℓ_2 loss function $f(U) = \sum_i (y_i - a_i^T UU^T a_i)^2$; we also analyze the quadratic nonnegative matrix factorization (QNMF) approach to clustering where the matrix to be factorized is the transition matrix for a reversible Markov chain.

In all of these instances, the optimization problem we wish to solve has many local optima and is highly non-convex. Instead of analyzing convex relaxations, which tend to be complicated and computationally expensive, we operate directly on the natural non-convex problems and prove both local and global optimality guarantees for a family of algorithms.

Table of Contents

Chapter 1

Introduction and Background

1.1 Background

Low rank matrices form the basis for many popular techniques in scientific computing and machine learning; recall that every rectangular matrix $M \in \mathbb{R}^{n \times p}$ has a singular value decomposition (SVD) given by

$$M = U_{n \times r} \Sigma_{r \times r} V_{n \times r}^T$$

where Σ is a diagonal matrix with positive entries and r is the *rank* of the matrix. Whenever $r \ll \min(n, p)$ it is clear that there are redundancies between the rows, columns and entries of M that we can exploit for the purposes of storage, computation or exploration. For example, techniques such as dimensionality reduction [26, 45, 33, 37] exploit this low-dimensional structure to reduce the number of variables and produce enlightening visualizations of very high dimensional datasets such as word clouds. Other popular utilizations of low rank structure include spectral clustering [40, 76], non-negative matrix factorization [49, 72, 56, 66], recommendation systems [21, 17], and many more. In all of these instances, we receive data that was generated from some process involving an unknown low-rank matrix and we seek to recover the low-rank matrix structure efficiently.

Solving an optimization problem with a matrix rank constraint is inherently non-convex. This is easy to see by noting that we can add two low rank matrices and potentially *double* the rank, thus violating the constraint. Inspired by exact recovery guarantees developed in the field of Compressed Sensing [19, 34], for a large class of low rank problems a great deal of research and attention was devoted to finding and analyzing appropriate convex relaxations [24, 67, 17, 18, 2]. In all of these works, randomness is leveraged to show that *with high probability* the solutions to NP-hard problems can be found by solving a simpler (but possibly larger) *convex* problem. The classical example of this is finding the sparsest vector satisfying a set of linear constraints, which can typically be solved via ℓ_1 minimization.

In this thesis, we will concern ourselves with the following situation: suppose we have a class of functions parameterized by a collection of data and we wish to solve

$$\begin{aligned} &\min_{M} f(M) \\ &\text{s.t. } M \in \mathcal{A} \end{aligned} \tag{1.1}$$

where the function $f(\cdot)$ is a possibly convex function, but the constraint set $\mathcal{A}$ is not. In particular, all of the problems considered in this thesis will have the constraint

$$\text{rank}(M) = r \ll n \tag{1.2}$$

where n is some ambient dimensionality. Typical methods for solving (1.1) involve relaxing the rank constraint (1.2) to the convex constraint

$$\|M\|_* \leq \eta$$

where $\|\cdot\|_*$ is the nuclear norm of M. This is directly analogous to ℓ_1 minimization for recovering sparse vectors, as sparsity of singular values is equivalent to being low rank.

Here, we will instead directly impose (1.2) via a quadratic factorization

$$M = XY^T$$

where $X, Y \in \mathbb{R}^{n\times r}$ are tall skinny matrices, possibly required to have orthogonal columns. It is clear that this transformation produces a large set of optimal solutions given by (XO, YO) for any orthogonal matrix $O \in \mathcal{O}(r)$. With this reparametrization, we will then carefully analyze algorithms which operate directly on the problem (1.1) and in many instances prove global optimality guarantees with high probability. A description of the main results are included in the next section.

1.2 Main Results

1.2.1 Quadratic Sampling

Consider $X \in \mathbb{R}^{n\times r}$ fixed and unknown, acquired through quadratic measurements of the form $y_i := \mathrm{tr}(XX^T a_i a_i^T) = \|X^T a_i\|_2^2$, $i = 1 \dots m$. Assuming the columns of X are orthogonal, this is equivalent to receiving noiseless samples $a_i^T M a_i$ of the positive semidefinite matrix $M := XX^T$. The question arises: given the data $\{(y_i, a_i)\}_{i=1}^m$, can we recover any information about the underlying matrix X?

Scenarios such as this arise in various applications: for a concrete ex-

ample, suppose we receive a stream of high-dimensional centered Gaussian vectors $\{x_j\}_{j=1}^m$ with unknown covariance matrix Σ. If we believe Σ to be (approximately) low rank, we can recover this structure via the sampled matrix

$$\hat{\Sigma} = \frac{1}{m}\sum_j x_j x_j^T.$$

However, due to the large dimensionality, storing all of the incoming vectors x_i might be prohibitive. Instead, we randomly draw a set of sensing vectors $\{a_k\}$ which are *sparse*, and for each incoming data point compute $y_{ki} = (a_k^T x_i)^2$. We are now only storing $(y_{ki}, a_k)_{k,i}$ which is a sparse data set. Note that if we define

$$y_k := \frac{1}{m}\sum_i y_{ki}$$

then y_k has the form

$$a_k^T(\frac{1}{m}\sum_j x_j x_j^T)a_k = a_k^T\hat{\Sigma}a_k.$$

The question posed above is: can we compute the covariance structure of the $\{x_i\}$ given only this data?

The example above describes *covariance sketching* of high-dimensional data streams [28, 24], but there are many other scenarios that fall under our problem setting, e.g., phaseless measurements in physics and optics [65, 71, 44, 39]. Because this data is invariant under the transformation

$$X \mapsto XO \tag{1.3}$$

for any orthogonal matrix $O \in \mathbb{R}^{r\times r}$, we can only hope to recover X up to this action. In the complex rank one setting $x \in \mathbb{C}^n$, this means we can only

recover x up to a global phase. *Phase retrieval* problems of this type often arise in the physical sciences due to the nature of optical sensors, which can only record intensity information [44, 39]. It is now well-understood that if the measurement vectors a_i are generic or e.g. independent Gaussian random vectors, then $m \geq O(n)$ measurements suffice for injectivity of the map $x \mapsto (|\langle a_i, x\rangle|^2)_{i=1}^m$ up to phase [5, 35]. There is still a question of how to perform the inverse map in an efficient and stable manner, and in recent years several different algorithms have been proposed in this direction, see for example [4, 3, 61, 18, 16, 1, 31, 35, 15]. In particular, [4] noted that such measurements may be reformulated as $y_i = \mathrm{tr}\left(a_i a_i^* x x^*\right)$, so that one can consider this problem as that of recovering an unknown rank-one positive semi-definite matrix. Inspired by the field of compressive sensing [19, 34] and low-rank matrix recovery [67], this led to many results demonstrating that well chosen convex and semidefinite programming (SDP) relaxations can provably recover the underlying signal up to phase with only $m \geq Cn$ Gaussian measurements [20, 18, 78]. However, as such algorithms optimize over the "lifted" space of $n \times n$ positive semidefinite matrices, the computational complexity becomes quite high. In the more general rank-r setting, whereby the measurements are $y_i := \mathrm{tr}(XX^T a_i a_i^T) = \|X^T a_i\|_2^2$, the recent works [54, 24] demonstrate that convex relaxation techniques based on nuclear norm minimization can solve such problems from an optimal number of measurements $O(nr)$, but still require large computational cost.

In the rank-1 setting in particular, several alternative reconstruction

algorithms have been proposed with global phase recovery guarantees which operate directly on the lower-dimensional problem, and thus are more computationally efficient. Notably, [61] considers the nonconvex optimization problem

$$\min_{u \in \mathbb{C}^n} \frac{1}{4m} \sum_{i=1}^{m} (y_i - |a_i^* u|^2)^2 \tag{1.4}$$

and proves that after a judiciously chosen initialization, with high probability alternating minimization will converge to the underlying vector x up to phase, assuming random Gaussian measurements. Subsequently [14] used the same initialization to show convergence when followed by gradient descent without requiring resampling. Both of these algorithms provably recover the underlying vector x up to global phase, from a number of measurements m which is optimal up to additional logarithmic factors in n. Very recently, the paper [23] provides a modified gradient method which removes the additional logarithmic factors of n in the number of measurements.

In a similar vein, many recent works have demonstrated global convergence guarantees for gradient descent on other nonconvex matrix factorization problems. Specifically, in [82] the authors consider gradient descent on the Grassmannian and prove global convergence for a class of SVD problems. In [30] a stochastic gradient algorithm was shown to converge globally for a low-rank matrix least squares problem. In all of these works including ours, the underlying idea is that the lack of convexity can be fixed by operating on an appropriate matrix manifold.

In this thesis, we consider the more general version of problem (1.4) in

which the underlying matrix is of rank r:

$$\min_{U \in \mathbb{R}^{n \times r}} f(U) := \min_{U \in \mathbb{R}^{n \times r}} \frac{1}{4m} \sum_{i=1}^{m} (y_i - \|a_i^T U\|_2^2)^2 \tag{1.5}$$

and our measurements take the form

$$y_i := \mathrm{tr}(XX^T a_i a_i^T) = \|X^T a_i\|_2^2. \tag{1.6}$$

Instead of constructing convex relaxations, we operate directly on the non-convex problem (1.5). Moreover, we demonstrate that under a Gaussian assumption on the random measurement vectors the function appearing in (1.5) is *strongly convex* in directions orthogonal to the manifold of solutions, and we can recover X (up to (1.3)) via spectral initialization followed by gradient descent.

Specifically, we have the following main theorem:

Theorem 1.2.1 (Main Theorem)**.** *Suppose we take* $m \geq C\|X\|_F^8 \lambda_r^{-4} n r^2 (\log n)^2$ *samples of the form* (1.6)*, where*

$$\lambda_1 \geq \lambda_2 ... \geq \lambda_r > 0$$

are the ordered eigenvalues of the matrix XX^T*. Define the matrix*

$$M := \frac{1}{2m} \sum_{i=1}^{m} y_i a_i a_i^T$$

and the following associated quantities:

$$\begin{aligned} U &:= \begin{bmatrix} u_1 & u_2 & ... & u_r \end{bmatrix}_{n \times r} \\ \Sigma &:= \begin{bmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & ... & 0 \\ ... & ... & 0 & \sigma_r \end{bmatrix}_{r \times r} - \sigma_{r+1} Id \\ U_0 &:= U \Sigma^{1/2} \end{aligned}$$

where $\sigma_1 \geq \sigma_2 ... \geq \sigma_{r+1} > 0$ *are the eigenvalues of* M *and* u_i *are the corresponding normalized eigenvectors. If we iteratively update* U_k *via gradient descent on* (1.5)

$$U_{k+1} = U_k - \gamma \nabla f(U)$$

for some constant step size

$$\gamma < C(\lambda_r/\lambda_1)^2 (nr)^{-4}$$

then with probability at least $1 - 3e^{-rn} - 7/m^2$ *we have that*

$$d(U_k) \leq \left[1 - 2\gamma\ell + \gamma^2 L^2\right]^k d(U_0)$$

where

$$\ell = \frac{\lambda_r^2}{18}$$
$$L = Cn^2 r^2 \lambda_1.$$

We believe this method of analysis should find use in providing recovery guarantees by gradient descent for a much broader class of non-convex problems arising in machine learning applications such as matrix completion, nonnegative matrix factorization, clustering, etc. More generally, such an analysis could possibly be useful towards achieving provable guarantees for machine learning problems which have many unstable saddle points, such as neural networks [29].

There are two main ingredients to proving Theorem 1.2.1, namely, the strong convexity of the function f in a region around the manifold of global

minimizers at finite sample complexity, and a guarantee that spectral initialization will land within this region. For details and the proof, see 2.1.1. In fact, the finite sample convexity result holds in more generality when $r = 1$, while for general r we always assume Gaussian measurements. See 2.2 for details on the rank one case.

1.2.2 Quadratic NMF

In many scenarios in data analysis, we are given a *nonnegative* matrix whose entries represent some notion of similarity between data points. A classical example is a rectangular word count matrix whose W_{ij} entry encodes the number of times vocabulary word j appears in document i. In this situation we can leverage nonnegativity to produce stronger models with powerful results; for example, Latent Semantic Indexing [55] was a breakthrough idea for language processing that essentially computes the SVD of the word count matrix. This model was improved upon with Probabilistic Latent Semantic Indexing [49] which uses nonnegativity in an essential way to produce a probabilistic model for the entries of the matrix. PLSI was further developed into a fully Bayesian model, Latent Dirichlet Allocation [7], which is perhaps the most well-known natural language processing tool.

Another important family of nonnegative matrices are square adjacency matrices whose entries represent similarities between data points, thought of as nodes in a weighted graph. The goal in this setting is typically to find *clusters* of data points that represent meaningful groups. For example, the so-called

k-medoids clustering problem can be set up as a binary integer programming problem [52] derived from the adjacency matrix; there are some non-convex algorithms for solving this problem (e.g., [42]) that are empirically successful but have no guarantees of convergence. Recently, in [2] it was shown that under certain conditions the globally optimal clustering can be found via a convex program.

A closely related family of nonnegative matrices are transition matrices for finite state Markov Chains, where P_{ij} represents the probability of a walker at state j moving to state i in the next step. There is a very rich and fascinating theory for such matrices that we will only touch on briefly in the proceeding chapters. In this setting, we might try to find approximately closed subsystems of states which have small probability of interacting with each other. This objective has connections to Simon-Ando theory [70] and stochastic complementation [60].

In this thesis, we will focus our attention on a particular nonnegative matrix factorization problem, whose motivation we defer to Chapter 3. The initial problem is as follows: given a symmetric nonnegative matrix S and an integer k, we seek to solve the non-convex optimization problem

$$\begin{aligned} &\min_{U \in \mathbb{R}^{n \times k}} \left\| S - UU^T \right\|_F^2 \\ &\text{s.t. } U_{ij} \geq 0, U^T U = Id_{k \times k} \end{aligned} \tag{1.9}$$

where $Id_{k \times k}$ is the identity matrix. Note that the non-negativity and orthogonality of U imply its columns have disjoint supports; consequently we can think of (1.9) as finding the optimal block structure in S. This objective was

considered in [80] where a heuristic algorithm was shown to have impressive empirical performance on clustering many different datasets. Surprisingly, in §3.1 we will derive (1.9) from a continuous problem arising in partial differential equations.

By expanding out the Frobenius norm, it is easy to see that (1.9) is equivalent to

$$\begin{aligned} &\max_{u_1,\ldots,u_k} \sum_{i=1}^{k} u_i^T S u_i \\ &\text{s.t. } u_i \geq 0, u_i^T u_j = \delta_{ij} \end{aligned} \tag{1.10}$$

which we recognize as an eigenvalue optimization problem in which we are attempting to find k disjoint principal submatrices of S such that the sum of their leading eigenvalues is maximal.

Leaving the symmetric world, we can then consider the more general problem of solving

$$\max \sum_{i=1}^{k} \rho(S_{\Omega_i}) \tag{1.11}$$

where the maximum is taken over all k-fold partitions of the nonnegative, irreducible matrix S, and $\rho(S_{\Omega_i})$ denotes the spectral radius of the principal submatrix formed from the indices in i. This is well-defined by standard results in Perron-Frobenius theory.

In Chapter 3, we will consider all of the above problems (1.9), (1.10), (1.11), providing rigorous motivations for each; we will then establish convergence guarantees for a novel *rearrangement algorithm* and an appropriate scaling limit of this algorithm. Connections to subspace clustering [64, 75] and

EM algorithms will be explored, leaving many intriguing directions for future research.

Chapter 2

Low-Rank Recovery from Quadratic Samples

Recall that we aim to solve the non-convex inverse problem of recovering an unknown matrix $X \in \mathbb{R}^{n\times r}$ with orthogonal columns from quadratic measurements of the form

$$y_i := \|a_i^T X\|_2^2 \tag{2.1}$$

by solving the non-convex optimization problem

$$\min_{U\in\mathbb{R}^{n\times r}} f(U) := \min_{U\in\mathbb{R}^{n\times r}} \frac{1}{4m}\sum_{i=1}^{m}(y_i - \|a_i^T U\|_2^2)^2. \tag{2.2}$$

Because the function appearing in (2.2) is invariant under right multiplication by an orthogonal matrix, there is an entire manifold of solutions given by $\{XO : O \in \mathcal{O}(r)\}$ where $\mathcal{O}(r)$ is the set of $r \times r$ orthogonal matrices. Our strategy is to establish that a spectral initialization will land (with high probability) in a region of strong convexity[1] around the manifold of global minimizers. An overview of our approach is given in Algorithm 1.

There are two main ingredients to proving performance guarantees for Algorithm 1, namely, the strong convexity of the function f in a region around

[1]Strong convexity here and throughout this chapter always refers to convexity in directions orthogonal to the manifold of solutions.

Algorithm 1 Initialize and Descend for finding global solutions to (2.2).

Input: Measurements $y_i = \|a_i^T U\|^2, \quad i = 1, 2, \ldots, m$, where a_i are Gaussian.

Initialize: $U_0 = Z\Lambda^{1/2}$, where the columns of Z contain the (ℓ_2-normalized) eigenvectors corresponding to the r largest eigenvalues $\sigma_1 \geq ... \geq \sigma_r$ of the matrix

$$M := \frac{1}{2m} \sum_{i=1}^{m} y_i a_i a_i^T$$

and the scaling is given by the diagonal matrix

$$\Lambda_i := \sigma_i - \sigma_{r+1}.$$

Descend: Starting at U_0, iteratively update U via gradient descent on $f(U)$.

Output: Estimated global solution $\widehat{X}$ to the nonconvex problem (2.39).

the manifold of global minimizers at finite sample complexity, and a guarantee that spectral initialization will land within this region. As previously mentioned, the finite sample convexity result holds in more generality when $r = 1$, while for general r we always assume Gaussian measurements.

The rest of this chapter is organized as follows: in §2.1.1 we prove the main finite sample convexity result, which relies on classifying tangent and normal directions to the manifold of solutions $\{XO : O \in \mathcal{O}(r)\}$ and an explicit formula for the expected Hessian. In §2.1.2 we prove that with high probability the initialization step produces a matrix in a convex region around the manifold of solutions and establish the convergence of gradient descent. In §2.2.1 we prove convexity results for the rank one case under more general

randomness assumptions, and in §2.2.2 we discuss the issues that arise with initialization under more general measurement schemes.

2.1 Rank r

2.1.1 Convexity

In the rank r setting, we have an entire manifold of solutions given by $\{XO : O \in \mathcal{O}(r)\}$; consequently we will need to consider the quantity

$$d(U) := \min_{O \in \mathcal{O}(r)} \|XO - U\|_F^2 \tag{2.3}$$

which is well-defined by compactness of the orthogonal group. We note that the minimizer may not be unique, but this is not important for our purposes. We will also need to consider

$$\lambda_1 \geq \lambda_2 ... \geq \lambda_r > 0$$

the non-zero eigenvalues of the positive semidefinite matrix XX^T.

The main lemma we rely on is the following simple characterization of the normal directions to the manifold of solutions:

Lemma 2.1.1. *Assume X has orthogonal columns and let*

$$O^* = \arg\min_{O \in \mathcal{O}(r)} \|XO - U\|_F^2$$

which is not necessarily unique. Then we can write

$$UO^{*^T} = X(X^TX)^{-1}M + P_\perp U$$

where $M \in \mathbb{R}^{r\times r}$ is a symmetric positive semidefinite matrix and $P_\perp$ is the projection onto the orthogonal complement of the column space of X.

Proof. This basically follows from the solution to the Orthogonal Procrustes Problem [69]. If we write $X^TU = ZDV^T$ for the singular value decomposition of X^TU, then we can expand the objective as follows:

$$\begin{aligned}
\arg\min_{O\in\mathcal{O}(r)} \|XO-U\|_F^2 &= \arg\min_{O\in\mathcal{O}(r)} \|X\|_F^2 + \|U\|_F^2 - 2\langle XO, U\rangle_F \\
&= \arg\max_{O\in\mathcal{O}(r)} \langle XO, U\rangle_F \\
&= \arg\max_{O\in\mathcal{O}(r)} \langle O, ZDV^T\rangle_F \\
&= Z\left(\arg\max_{O'\in\mathcal{O}(r)} \langle O', D\rangle_F\right) V^T \\
&= ZV^T.
\end{aligned}$$

We then find that

$$X^TUO^{*^T} = ZDZ^T$$

is a symmetric positive semidefinite matrix. As $X^TA = 0$ is equivalent to $P_\perp A = A$, we arrive at the stated claim. □

This lemma says that if we consider the direction $W = U - XO^*$ between U and its closest solution matrix XO^* we have that

$$O^{*^T}X^TW = O^{*^T}\left(M - X^TX\right)O^*$$

which is a *symmetric* matrix. Why symmetry is important will become apparent after the next lemma, which establishes formulas for the expectation of the Hessian of (2.2):

Lemma 2.1.2. *The gradient of* $f(U) = \frac{1}{4m}\sum_{i=1}^{m}(y_i - a_i^T UU^T a_i)^2$ *is given by*

$$\nabla f(U) = \begin{bmatrix} \nabla f_1(U) & \dots & \nabla f_r(U) \end{bmatrix} \in \mathbb{R}^{n\times r}$$

where

$$\nabla f_k(U) = \frac{1}{m}\sum_{i=1}^{m}(a_i^T UU^T a_i - y_i)(a_i^T u_k)a_i^T \tag{2.5}$$

and the Hessian of $f(U)$ *is given by*

$$\nabla^2 f(U) = \frac{1}{m}\sum_{i=1}^{m} H^{(i)} \tag{2.6}$$

where the $nr \times nr$ *matrices* $H^{(i)}$ *are given by*

$$\begin{aligned} H_{jk}^{(i)} &= \begin{bmatrix} 2(a_i^T u_j)(a_i^T u_k)a_i a_i^T \end{bmatrix} \\ H_{jj}^{(i)} &= \begin{bmatrix} (a_i^T UU^T a_i + 2(a_i^T u_j)^2 - y_i) \end{bmatrix} \end{aligned}$$

and if we suppose the a_i*'s are i.i.d. centered Gaussian random vectors satisfying* $\mathbb{E}[aa^T] = Id$*, then the expectation of* (2.6) *is given by*

$$\mathbb{E}[\nabla^2 f(X)] = A + D \tag{2.8}$$

where the $nr \times nr$ *block matrices* A *and* D *satisfy*

$$A_{ij} = \begin{bmatrix} 2u_i u_j^T + 2u_j u_i^T + 2(u_i^T u_j)Id \end{bmatrix}_{n\times n} \tag{2.9a}$$

$$D_{jj} = \begin{bmatrix} (\|U\|_F^2 - \|X\|_F^2)Id + 2(UU^T - XX^T) \end{bmatrix}_{n\times n}.$$

Proof. The proofs of (2.5) and (2.6) use standard vector calculus. For (2.8), we use the fact that for any vector $x \in \mathbb{R}^n$

$$\mathbb{E}[(x^T a)^2 aa^T] = \|x\|_2^2 Id + 2xx^T$$

which can be seen by writing out the entries of the matrix individually. This implies

$$\begin{aligned}\mathbb{E}[(a^T XX^T a)aa^T] &= \sum_{i=1}^{r} \mathbb{E}[(a^T x_i)^2 aa^T] \\ &= \sum_{i=1}^{r} \left(\|x_i\|_2^2 Id + 2x_i x_i^T\right)\end{aligned}$$

which, combined with

$$\mathbb{E}[(a^T x_i)(a^T x_j)aa^T] = x_i x_j^T + x_j x_i^T + x_i^T x_j Id$$

yields the stated result. □

Note that (2.9a) shows us that for a matrix $W = \begin{bmatrix} w_1 & ... & w_r \end{bmatrix}_{n\times r}$ we can write

$$\begin{aligned}\frac{1}{2}\text{vec}(W)^T \mathbb{E}[\nabla^2 f(X)]\text{vec}(W) &= \sum_{i=1}^{r} 2(x_i^T w_i)^2 + \lambda_i \|w_i\|_2^2 \\ &\quad + \sum_{i\neq j} (w_i^T x_i)(w_j^T x_j) + (w_i^T x_j)(w_j^T x_i) \\ &= \text{tr}\left(X^T W\right)^2 + \sum_{i=1}^{r} \lambda_i \|w_i\|_2^2 + \text{tr}\left(X^T W X^T W\right).\end{aligned}$$

Example 1. If $W = X(X^T X)^{-1} X^T S$ where $(X^T S) = -S^T X$ is an $r \times r$ skew-symmetric matrix, then we have

$$\frac{1}{2}\text{vec}(W)^T \mathbb{E}[\nabla^2 f(X)]\text{vec}(W) = \text{tr}\left(X^T S\right)^2 + \sum_{i=1}^{r} \lambda_i \|w_i\|_2^2 + \text{tr}\left((X^T S)^2\right)$$

Note that the diagonal of $X^T S$ is necessarily 0 implying $\text{tr}\left(X^T S\right)^2 = 0$ and it

is easy to see that skew symmetry implies

$$\begin{aligned}\operatorname{tr}\left((X^TS)^2\right) &= -\sum_{i=1}^{r}\|s_i\|_2^2 \\ &= -\sum_{i=1}^{r}\lambda_i\|w_i\|_2^2\end{aligned}$$

where s_i is the ith column of X^TS. Consequently the function $\mathbb{E}[f]$ is flat in the direction W.

Example 2. If $W = X(X^TX)^{-1}X^TM$ where $X^TM = M^TX$ is an $r \times r$ symmetric matrix, then we have

$$\frac{1}{2}\mathrm{vec}(W)^T\mathbb{E}[\nabla^2 f(X)]\mathrm{vec}(W) = \operatorname{tr}\left(X^TM\right)^2 + \sum_{i=1}^{r}\lambda_i\|w_i\|_2^2 + \operatorname{tr}\left((X^TM)^2\right).$$

In this case symmetry gives us

$$\begin{aligned}\operatorname{tr}\left((X^TM)^2\right) &= \left\|X^TM\right\|_F^2 \\ &\geq 0\end{aligned}$$

and

$$\sum_{i=1}^{r}\lambda_i\|w_i\|_2^2 \geq \lambda_r\|W\|_F^2 \tag{2.12}$$

so that the function $\mathbb{E}[f]$ is *strongly convex* in the direction W.

Before establishing the main theorem, we need a standard concentration result:

Lemma 2.1.3. *Suppose we collect $m \geq C\delta^{-2}\beta nr\log(n)$ samples of the form $y_i := a_i^TXX^Ta_i$, where δ and β are given constants and $r = \operatorname{rank}(X)$; then we*

have that with probability greater than $1 - 2e^{-\beta rn} - 6/m^2$

$$\left\|\nabla^2 f(X) - \mathbb{E}\left[\nabla^2 f(X)\right]\right\|_{op} < 2\delta\|X\|_{op}^2.$$

The sampling complexity can be improved, but we state Lemma 2.1.3 as a general proof-of-concept. For details and a proof see §2.1.3.

With the above Lemmas at hand, we can now lower bound the Hessian of (1.5) at a point U, via the following:

Theorem 2.1.4 (Strong Convexity)**.** *Suppose we take* $m \geq C(\lambda_r/\lambda_1)^{-2}nr(\log n)^2$ *samples of the form* (2.1) *and that* $U \in \mathbb{R}^{n\times r}$ *satisfies*

$$\min_{O\in\mathcal{O}(r)} \|XO - U\|_F < \frac{3\lambda_r}{10\|X\|_F}. \tag{2.13}$$

Then with probability at least $1 - 4e^{-rn} - 7/m^2$*, it holds that*

$$\begin{aligned}\mathrm{vec}(U - XO^*)^T\nabla^2 f(U)\mathrm{vec}(U - XO^*) &\geq \frac{\lambda_r^2}{18\|X\|_F^2}\\ \mathrm{vec}(U - XO^*)^T\nabla^2 f(U)\mathrm{vec}(U - XO^*) &\leq Cn^2r^2\lambda_1\end{aligned}$$

where O^* *is a minimizer for* (2.13)*.*

Note that this theorem implies that for matrices $U \in \mathbb{R}^{n\times r}$ close to the manifold of solutions, we can control the eigenvalues of the Hessian; in particular for such U, *the function* $f(U)$ *is strongly and uniformly convex on the line connecting* U *to its nearest point on the manifold of solutions*, as measured by the function (2.3).

We will rely on the following Lemma from [6], as stated in [14]:

Lemma 2.1.5. *Suppose $Y_1, Y_2, ..., Y_m$ are i.i.d. real-valued random variables obeying $Y_i \leq b$ for some nonrandom $b > 0$, $\mathbb{E}[Y_i] = 0$, and $\mathbb{E}[Y_r^2] = v^2$. Setting $\sigma^2 = m \cdot \max(b^2, v^2)$,*

$$\mathbb{P}\left[Y_1 + Y_2 + ... + Y_m \geq y\right] \leq \min\left\{\exp\left(-\frac{y^2}{\sigma^2}\right), c_0(1 - \Phi(y/\sigma))\right\}$$

where one can take $c_0 = 25$ and $\Phi(\cdot)$ is the CDF for the standard normal.

Proof of Theorem 2.1.4. Let $\hat{W} := W/\|W\|_F$ be the normalized direction from U to XO^* and let $t \geq 0$ be a positive scalar. Moreover, WLOG we will be assuming that $\|X\|_F = 1$. Note that $f(U)$ is invariant under the action of $\mathcal{O}(r)$ and thus it suffices to consider the case where $O^* = Id$.

By abuse of notation, consider the single-variable function $f(X + t\hat{W})$ which can be written

$$f(t) = \frac{1}{4m}\sum_{i=1}^{m}\left(2ta_i^T X\hat{W}^T a_i + t^2 a_i^T \hat{W}\hat{W}^T a_i\right)^2.$$

It is straightforward to verify that

$$f''(t) = \frac{1}{m}\sum_{i=1}^{m} 3\left(a_i^T\hat{W}\hat{W}^T a_i\right)^2 t^2 + 6\left(a_i^T\hat{W}\hat{W}^T a_i\right)\left(a_i^T X\hat{W}^T a_i\right)t + 2\left(a_i^T X\hat{W}^T a_i\right)^2 \tag{2.14}$$

which is a convex polynomial in t; observe that $f''(0) > 0$. If the linear term is positive, then clearly (2.14) is positive for all t and we have nothing to show (the smallest eigenvalue is bounded below by $f''(0)$ in the direction $\hat{W}$). Define the following quantities:

$$A_i := \left(a_i^T\hat{W}\hat{W}^T a_i\right) = \|\hat{W}^T a_i\|_2^2$$
$$B_i := \left(a_i^T X\hat{W}^T a_i\right) = \langle X^T a_i, \hat{W}^T a_i\rangle$$

and note that (2.14) can be written as

$$f''(t) = \frac{3}{m}\sum_{i=1}^{m}(A_i t + B_i)^2 - \frac{1}{m}\sum_{i=1}^{m} B_i^2.$$

Consider the random variable

$$Z_i(t) := (A_i t + B_i)^2 \geq 0$$

Observe that A_i is a chi-squared random variable with 1 degree of freedom by the normalization $\|\hat{W}\|_F = 1$. Thus we have

$$\begin{aligned}
\mathbb{E}[A_i] &= 1 \\
\mathbb{E}[A_i^2] &= 3 \\
\mathbb{E}[A_i^4] &= 105 \\
\mathbb{E}[A_i^6] &= 10395 \\
\mathbb{E}[A_i^8] &= 2027025 \\
\mathbb{E}[A_i^{12}] &= 18602008425.
\end{aligned}$$

By the definition of $f''(t)$, we have

$$f''(0) = \mathrm{vec}(\hat{W})^T \mathbb{E}\left[\nabla^2 f(X)\right] \mathrm{vec}(\hat{W})$$

and so

$$\mathbb{E}[B_i^2] = \frac{1}{2}\mathrm{vec}(\hat{W})^T \mathbb{E}\left[\nabla^2 f(X)\right] \mathrm{vec}(\hat{W}). \tag{2.17}$$

Moreover,

$$\begin{aligned}
\mathbb{E}[A_i B_i] &= \mathbb{E}\left[\left(\sum_{k=1}^{r}(a_i^T w_k)^2\right)\left(\sum_{k=1}^{r}(a_i^T x_k)(a_i^T w_k)\right)\right] \\
&= \mathbb{E}\left[\sum_{k,q=1}^{r}(a_i^T w_k)^2(a_i^T x_q)(a_i^T w_q)\right] \\
&= \sum_{k,q=1}^{r} 2(w_k^T x_q)(w_k^T w_q) + \|w_k\|_2^2 x_q^T w_q \\
&= 2\sum_{k,q=1}^{r} x_q^T w_k w_k^T w_q + \sum_{q=1}^{r} x_q^T w_q \\
&= 2\sum_{q=1}^{r} x_q^T \hat{W}\hat{W}^T w_q + \mathrm{tr}(X^T\hat{W}) \\
&= 2\mathrm{tr}(X^T\hat{W}\hat{W}^T\hat{W}) + \mathrm{tr}(X^T\hat{W})
\end{aligned}$$

We then have that the mean $\mu(t) := \mathbb{E}[Z_i(t)]$ is given by

$$3t^2 + \left(4\mathrm{tr}(X^T\hat{W}\hat{W}^T\hat{W}) + 2\mathrm{tr}(X^T\hat{W})\right)t + \frac{1}{2}\mathrm{vec}(\hat{W})^T\mathbb{E}\left[\nabla^2 f(X)\right]\mathrm{vec}(\hat{W}). \tag{2.18}$$

Next we consider the variance of $Z_i(t)$:

$$\begin{aligned}
\mathbb{E}[(Z_i(t) - \mu(t))^2] &\leq \mathbb{E}[Z_i^2(t)] \\
&= \mathbb{E}\left[A_i^4\right] t^4 + 4\mathbb{E}\left[A_i^3 B_i\right] t^3 + 6\mathbb{E}\left[A_i^2 B_i^2\right] t^2 + 4\mathbb{E}\left[A_i B_i^3\right] t + \mathbb{E}\left[B_i^4\right] \\
&= 105t^4 + 4\mathbb{E}\left[A_i^3 \langle X^T a_i, \hat{W}^T a_i \rangle\right] t^3 \\
&\quad + 6\mathbb{E}\left[A_i^2 \langle X^T a_i, \hat{W}^T a_i \rangle^2\right] t^2 + 4\mathbb{E}\left[A_i \langle X^T a_i, \hat{W}^T a_i \rangle^3\right] t \\
&\quad + \mathbb{E}\left[\langle X^T a_i, \hat{W}^T a_i \rangle^4\right] \\
&\leq 105t^4 + 4\sqrt{\mathbb{E}[A_i^6]\mathbb{E}\left[\langle X^T a_i, \hat{W}^T a_i \rangle^2\right]} t^3 \\
&\quad + 6\mathbb{E}[A_i^3 \|X^T a_i\|_2^2] t^2 + 4\sqrt{\mathbb{E}[A_i^2]\mathbb{E}[\langle X^T a_i, \hat{W}^T a_i \rangle^6]} t \\
&\quad + \mathbb{E}\left[\|X^T a_i\|_2^4 \|\hat{W}^T a_i\|_2^4\right]
\end{aligned}$$

where we used either Cauchy-Schwarz or Hölder's inequality on each term. Proceeding with applications of Hölder and Cauchy-Schwarz, recognizing that $\|X^T a_i\|_2^2$ is also Chi-squared with one degree of freedom under the assumption

that $\|X\|_F = 1$, and noting from (2.17) that $\mathbb{E}[B_i^2] \leq 3$ we have

$$
\begin{aligned}
&= 105t^4 + 4\sqrt{10395 \cdot \mathbb{E}\left[B_i^2\right]}t^3 \\
&\quad + 6\mathbb{E}[A_i^3\|X^T a_i\|_2^2]t^2 + 4\sqrt{3 \cdot \mathbb{E}[\langle X^T a_i, \hat{W}^T a_i\rangle^6]}t + \mathbb{E}\left[\|X^T a_i\|_2^4 \|\hat{W}^T a_i\|_2^4\right] \\
&\leq 105t^4 + 707t^3 + 6\sqrt{\mathbb{E}[A_i^6\|X^T a_i\|_2^4]}t^2 + 4\sqrt{3 \cdot \mathbb{E}[\|X^T a_i\|_2^6 \|\hat{W}^T a_i\|_2^6]}t \\
&\quad + \sqrt{\mathbb{E}\left[\|X^T a_i\|_2^8\right]\mathbb{E}[\|\hat{W}^T a_i\|_2^8]} \\
&\leq 105t^4 + 707t^3 + 6\left(\mathbb{E}[A_i^{12}]\mathbb{E}[\|X^T a_i\|_2^8]\right)^{1/4} t^2 \\
&\quad + 4\sqrt{3}\left(\mathbb{E}[\|X^T a_i\|_2^{12}]\mathbb{E}[\|\hat{W}^T a_i\|_2^{12}]\right)^{1/4} t + 105 \\
&\leq 105t^4 + 707t^3 + 7094t^2 + 707t + 105 \\
&= C(t)^2.
\end{aligned}
$$

Observe that the mean can be bounded as

$$\mu(t) \leq 3t^2 + 6t + 3\lambda_1.$$

Now define

$$
\begin{aligned}
Y_i(t) &:= \mu(t) - Z_i(t) \\
b(t) &:= 3t^2 + 6t + 3\lambda_1 \\
v^2(t) &:= C(t)^2 \\
\sigma^2(t) &:= mC(t)^2 \\
y &:= m\lambda_r/12.
\end{aligned}
$$

Applying Lemma 2.1.5 above yields

$$\mathbb{P}\left[\mu(t) - \frac{1}{m}\sum_{i=1}^{m} Z_i(t) \geq \lambda_r/12\right] \leq \min\left\{\exp\left(-\frac{\lambda_r^2 m}{144C(t)^2}\right), 25\left(1-\Phi(\frac{\lambda_r\sqrt{m}}{12C(t)})\right)\right\}.$$

Using the well-known bound

$$1-\Phi(\frac{\lambda_r\sqrt{m}}{12C(t)}) < \frac{12C(t)}{\lambda_r\sqrt{2m\pi}}\exp\left(-\frac{\lambda_r^2 m}{288C(t)^2}\right)$$

we find that if $m \geq 288\alpha\lambda_r^{-2}C(t)^2 nr$ then with probability at least $1-e^{-\alpha nr}$ we have

$$\begin{aligned} f''(t) &= \frac{3}{m}\sum_{i=1}^{m}(A_i t + B_i)^2 - \frac{1}{m}\sum_{i=1}^{m} B_i^2 \\ &\geq 3\mu(t) - \lambda_r/4 - \frac{1}{m}\sum_{i=1}^{m} B_i^2 \\ &\geq 2t^2 - 6t - \lambda_r/4 + \frac{3}{2}\mathrm{vec}(\hat{W})^T\mathbb{E}\left[\nabla^2 f(X)\right]\mathrm{vec}(\hat{W}) - \frac{1}{2}\mathrm{vec}(\hat{W})^T\nabla^2 f(X)\mathrm{vec}(\hat{W}) \end{aligned} \tag{2.20}$$

where we used (2.18) to lower bound $\mu(t)$ by

$$t^2 - 6t + \frac{1}{2}\mathrm{vec}(\hat{W})^T\mathbb{E}\left[\nabla^2 f(X)\right]\mathrm{vec}(\hat{W})$$

and the fact that

$$\frac{1}{m}\sum_{i=1}^{m} B_i^2 = \frac{1}{2}\mathrm{vec}(\hat{W})^T\nabla^2 f(X)\mathrm{vec}(\hat{W}).$$

Moreover, an ϵ-net argument over all directions $\hat{W}$ shows that (2.20) holds for an arbitrary $\hat{W}$ with probability at least $1-e^{-\beta nr}$.

Further observe that our condition on m guarantees

$$\left\|\nabla^2 f(X) - \mathbb{E}\left[\nabla^2 f(X)\right]\right\|_{op} < \frac{\lambda_r}{4}$$

with probability at least $1 - 2e^{-\beta rn} - 6/m^2$ by Lemma 2.1.3. By (2.12) this implies that

$$\frac{3}{2}\text{vec}(\hat{W})^T \mathbb{E}\left[\nabla^2 f(X)\right] \text{vec}(\hat{W}) - \frac{1}{2}\text{vec}(\hat{W})^T \nabla^2 f(X) \text{vec}(\hat{W}) > \frac{15}{8}\lambda_r$$

so that

$$f''(t) \geq 2t^2 - 6t + \frac{15}{8}\lambda_r$$

with probability at least $1 - 3e^{-\beta rn} - 6/m^2$ for any direction $\hat{W}$. Thus by a tangent line bound we find that the smallest positive root of $f''(t)$ is bounded below by

$$t^* \geq \frac{15}{48}\lambda_r > \frac{3}{10}\lambda_r$$

and we note that

$$\begin{aligned} f''(\frac{15}{48}\lambda_r) &= 2(\frac{15}{48})^2\lambda_r^2 + \frac{15}{8}(\lambda_r - \lambda_r) \\ &\geq \frac{\lambda_r^2}{18}. \end{aligned}$$

Thus $f''(t) \geq \frac{\lambda_r^2}{18}$ for all $t \in [0, \frac{3}{10}\lambda_r]$ which is the advertised lower bound.

For the upper bound, observe that by Cauchy-Schwarz

$$\frac{6}{m}\sum_{i=1}^{m}\left(a_i^T \hat{W}\hat{W}^T a_i\right)\left(a_i^T X \hat{W}^T a_i\right) \leq 6\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(a_i^T \hat{W}\hat{W}^T a_i\right)^2}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(a_i^T X \hat{W}^T a_i\right)^2}$$

and thus we find an upper bound for (2.14) is given by

$$f''(t) \leq 3a^2t^2 + 6abt + 2b^2 \tag{2.22}$$

where

$$a = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(a_i^T\hat{W}\hat{W}^T a_i\right)^2}$$
$$b = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(a_i^T X\hat{W}^T a_i\right)^2}.$$

Now, $\omega_i := \left(a_i^T\hat{W}\hat{W}^T a_i\right)$ is a chi-squared random variable with one degree of freedom; consequently we find

$$\mathbb{P}[\frac{1}{m}\sum_{i=1}^{m}\omega_i^2 \geq C^2n^2r^2] \leq \frac{m}{\sqrt{Cnr}}e^{-Cnr} \leq e^{-\tilde{C}nr}$$

as long as $m \geq Cnr$. Consequently an ϵ-net argument shows us that for *any* $\hat{W} \in \mathbb{R}^{n\times r},\ \|\hat{W}\|_F = 1$ we have

$$\frac{1}{m}\sum_{i=1}^{m}\left(a_i^T\hat{W}\hat{W}^T a_i\right)^2 \leq C^2n^2r^2$$

with probability greater than $1 - e^{-\beta nr}$.

Returning to (2.22) we find then that

$$\begin{aligned} f''(t) &\leq 3\left(at+b\right)^2 - b^2 \\ &\leq 3\left(Cnr\lambda_r + b\right)^2 \end{aligned}$$

for all $t \in [0, \frac{3}{10}\lambda_r]$.

To finish the proof, observe that we can write

$$b = \frac{1}{\sqrt{2}}\sqrt{\mathrm{vec}(\hat{W})^T\nabla^2 f(X)\mathrm{vec}(\hat{W})}$$

and by Lemma 2.1.3 we have

$$\begin{aligned} b &\leq \frac{1}{\sqrt{2}}\sqrt{\operatorname{vec}(\hat{W})^T \mathbb{E}\left[\nabla^2 f(X)\right] \operatorname{vec}(\hat{W}) + 2\delta\lambda_1} \\ &\leq \frac{1}{\sqrt{2}}\sqrt{6\lambda_1 + 2\delta\lambda_1} \end{aligned}$$

yielding

$$\begin{aligned} f''(t) &\leq 3\left(Cnr\lambda_r + \frac{1}{\sqrt{2}}\sqrt{6\lambda_1 + 2\delta\lambda_1}\right)^2 \\ &\leq Cn^2r^2\lambda_1. \end{aligned}$$

□

2.1.2 Initialization and Gradient Descent

Now that we have established the main strong convexity result in Theorem 2.1.4, it remains to certify a point within this region to initialize gradient descent. Recall the definition of the distance function (2.3), which is needed below.

We will also need a slightly different concentration result than previously, whose proof we defer to §2.1.3:

Lemma 2.1.6. *Suppose we collect $m \geq C\delta^{-2}\beta nr\log(n)$ samples of the form (1.6), where δ and β are given constants and $r = \operatorname{rank}(X)$; then we have that with probability greater than $1 - 2e^{-\beta rn} - 6/m^2$*

$$\left\| \frac{1}{m}\sum_{i=1}^{m} y_i a_i a_i^T - \|X\|_F^2 Id - 2X^* X^{*^T} \right\|_{op} < \delta\|X\|_F^2.$$

Theorem 2.1.7. *Suppose we take $m \geq C\beta\lambda_r^{-4}\|X\|_F^8 nr^2(\log n)^2$ samples. Define the matrix*

$$M := \frac{1}{2m}\sum_{i=1}^{m} y_i a_i a_i^T$$

and the following quantities:

$$\begin{aligned} U &:= \begin{bmatrix} u_1 & u_2 & ... & u_r \end{bmatrix}_{n\times r} \\ \Sigma &:= \begin{bmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & ... & 0 \\ ... & ... & 0 & \sigma_r \end{bmatrix}_{r\times r} - \sigma_{r+1} Id \\ U_0 &:= U\Sigma^{1/2} \end{aligned}$$

where $\sigma_1 \geq \sigma_2 ... \geq \sigma_{r+1} > 0$ are the eigenvalues of M and u_i are the corresponding normalized eigenvectors. Then with probability at least $1 - 3e^{-\beta rn} - 7/m^2$ we have that

$$d(U_0) < \frac{9}{100}\lambda_r^2 \tag{2.26}$$

where

$$d(U) := \min_{O\in\mathcal{O}_r} \|U - XO\|_F^2.$$

Proof of Theorem 2.1.7. It suffices to prove the case $\|X\|_F^2 = 1$. By Lemma 2.1.6 we have that with probability greater than $1 - 2e^{-\beta rn} - 6/m^2$

$$\left\| \frac{1}{m}\sum_{i=1}^{m} y_i a_i a_i^T - Id - 2XX^T \right\|_{op} < \delta$$

so long as $m \geq C\delta^{-2}\beta nr\log(n)^2$. This implies

$$\left\| M - (1/2)Id - XX^T \right\|_{op} < \delta/2.$$

Let

$$A := M - (1/2)Id$$

and collect the unit normalized eigenvectors corresponding to the dominant r-dimensional subspace of A in a matrix $U \in \mathbb{R}^{n\times r}$. Let $\sigma_1 \geq \sigma_2 ... \geq \sigma_{r+1} > 0$ denote the eigenvalues of the observed matrix M and define

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & ... & 0 \\ ... & ... & 0 & \sigma_r \end{bmatrix}_{r\times r} - \sigma_{r+1} Id_{r\times r}.$$

Let $U_0 := U\Sigma^{1/2}$, and $Q := O_1 O_2^T$ where $(X^TX)^{-1/2}X^TU = O_1 D O_2^T$ is the singular value decomposition and observe

$$\begin{aligned} \left\|U\Sigma^{1/2} - XQ\right\|_F &= \left\|U\Sigma^{1/2} - U(X^TX)^{1/2} + U(X^TX)^{1/2} - XQ\right\|_F \\ &\leq \left\|U - X(X^TX)^{-1/2}Q\right\|_F \left\|Q^T(X^TX)^{1/2}\right\|_{op} \\ &\quad + \|U\|_{op} \left\|\Sigma^{1/2} - (X^TX)^{1/2}\right\|_F \\ &\leq \frac{2^{3/2}\sqrt{r}\|A - XX^T\|_{op}}{\lambda_r}\sqrt{\lambda_1} + \left\|\Sigma^{1/2} - (X^TX)^{1/2}\right\|_F \qquad (2.27a) \\ &\leq \frac{2^{1/2}\sqrt{r}\delta}{\lambda_r}\sqrt{\lambda_1} + \frac{1}{2\lambda_r}\left(\sqrt{r}\delta + \delta\right) \qquad (2.27b) \end{aligned}$$

where (2.27a) follows from Theorem 2 in [81] and (2.27b) follows from

$$\begin{aligned}
\left\|\Sigma^{1/2} - (X^T X)^{1/2}\right\|_F &= \sqrt{\sum_{i=1}^{r} \left|\sqrt{\sigma_i - \sigma_{r+1}} - \sqrt{\lambda_i}\right|^2} \\
&= \sqrt{\sum_{i=1}^{r} \left|\frac{\sigma_i - \sigma_{r+1} - \lambda_i}{\sqrt{\sigma_i - \sigma_{r+1}} + \sqrt{\lambda_i}}\right|^2} \\
&\leq \frac{1}{\sqrt{\lambda_r}} \left\|M - XX^T\right\|_F \\
&\leq \frac{1}{\sqrt{\lambda_r}} \left(\left\|A - XX^T\right\|_F + |\sigma_{r+1} - 1/2|\right) \\
&\leq \frac{1}{2\lambda_r} \left(\sqrt{r}\delta + \delta\right).
\end{aligned}$$

The first equality holds because X has orthogonal columns and thus $X^T X$ is a diagonal matrix.

Now, we have

$$\begin{aligned}
\frac{2^{1/2}\sqrt{r}\delta}{\lambda_r}\sqrt{\lambda_1} &< \frac{3}{20}\lambda_r \\
\frac{1}{2\lambda_r}\left(\sqrt{r}\delta + \delta\right) &< \frac{3}{20}\lambda_r
\end{aligned}$$

so long as

$$\delta < C(\sqrt{r})^{-1} \min(\lambda_r^2/\sqrt{\lambda_1}, \lambda_r^2) = Cr^{-1/2}\lambda_r^2$$

which gives the stated claim. □

Initializing from a matrix satisfying (2.26) guarantees we are close enough so that gradient descent will converge, as established in the following:

Theorem 2.1.8. *Suppose we take $m \geq C\beta\lambda_r^{-4}\|X\|_F^8 nr(\log n)$ samples. Define the function*

$$d(U) := \min_{O \in \mathcal{O}(r)} \|XO - U\|_F^2$$

and suppose that $U_0 \in \mathbb{R}^{n\times r}$ satisfies

$$d(U_0) < \frac{9}{100}\lambda_r^2.$$

Then with probability at least $1 - 3e^{-\beta rn} - 7/m^2$ we have that

$$d(U_k) \leq \left[1 - 2\gamma\ell + \gamma^2 L^2\right]^k d(U_0)$$

where

$$\ell = \frac{\lambda_r^2}{18}$$
$$L = Cn^2r^2\lambda_1$$
$$U_{k+1} := U_k - \gamma\nabla f(U_k)$$
$$\gamma < \frac{2\ell}{L^2}.$$

Proof. Let $U^+ := U - \gamma\nabla f(U)$ and $O^* = \arg\min_{O\in\mathcal{O}(r)} \|XO - U\|_F^2$. Then we have

$$\begin{aligned}
d(U^+) &\leq \|U^+ - XO^*\|_F^2 \\
&= \|U - \gamma\nabla f(U) - XO^*\|_F^2 \\
&= \|U - XO^*\|_F^2 - 2\gamma\langle\nabla f(U), U - XO^*\rangle + \gamma^2\|\nabla f(U) - \nabla f(X)\|_F^2 \\
&\leq \|U - XO^*\|_F^2 - 2\gamma\ell\|U - XO^*\|_F^2 + \gamma^2 L^2\|U - XO^*\|_F^2 \qquad (2.31a) \\
&= \left[1 - 2\gamma\ell + \gamma^2 L^2\right]\|U - XO^*\|_F^2 \\
&= \left[1 - 2\gamma\ell + \gamma^2 L^2\right] d(U)
\end{aligned}$$

where for (2.31a) we used the convexity guaranteed by Theorem 2.1.4. □

A few remarks are in order.

1. Note that the quantity $\|X\|_F^8 \lambda_r^{-4}$ is scale invariant; however, we have the bounds
$$r^4 \leq \|X\|_F^8 \lambda_r^{-4} \leq r^4(\lambda_1/\lambda_r)^4.$$

2. One consequence of our result is that the sampling complexity is entirely independent of the desired solution tolerance. That is, the fixed set of $m \geq Cnr^6(\log n)^2$ samples suffices to produce a global solution up to arbitrary accuracy.

3. Our numerical results in §4.1 suggest that in general the sampling complexity only linearly depends on the ambient dimension n. Consequently a more refined analysis and initialization procedure such as that found in the recent work of [23] for the case of rank-1 recovery is most likely possible also in the general rank-r recovery setting.

4. This method of analysis should find use in providing recovery guarantees by gradient descent for a broader class of nonconvex problems arising in machine learning applications such as matrix completion, nonnegative matrix factorization, clustering, etc. More generally, such an analysis could possibly be useful towards achieving provable guarantees for machine learning problems which have many unstable saddle points, such as neural networks [29].

2.1.3 Concentration Results

In this section we collect the proofs of the concentration results needed above. We begin with the main concentration theorem:

Theorem 2.1.9. *Let $X \in \mathbb{R}^{n\times r}$ be a given matrix with orthogonal columns; suppose $m \geq C\delta^{-2}\beta nr\log(n)$ where δ and β are given constants and $r =$rank(X). Then we have that with probability greater than $1 - 2e^{-\beta rn} - 6/m^2$*

$$\left\| \frac{1}{m}\sum_{i=1}^{m} a_i a_i^T \otimes a_i a_i^T - \mathbb{E}[a_i a_i^T \otimes a_i a_i^T] \right\|_{\big|_{X\otimes\mathbb{R}^n}} < \delta$$

where $\|\cdot\|_{\big|_{X\otimes\mathbb{R}^n}}$ is the operator norm of the matrix restricted to the subspace spanned by the columns of X tensored with $\mathbb{R}^n$.

Proof. Let $x \otimes z \in X \otimes \mathbb{R}^n$ be an arbitrary unit vector. Write $z = z_\perp + w$, where $w \in X$ and $\langle z_\perp, x\rangle = 0$. We must consider the quantity

$$\begin{aligned}
&\Big|\frac{1}{m}\sum_{i=1}^{m}(a_i^T x)^2(a_i^T z)^2 - 2(x^T z)^2 - 1\Big| \\
&\quad = \Big|\frac{1}{m}\sum_{i=1}^{m}(a_i^T x)^2(a_i^T z_\perp)^2 - \|z_\perp\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}(a_i^T x)^2(a_i^T w)^2 - 2(x^T w)^2 - \|w\|_2^2 \\
&\qquad + \frac{2}{m}\sum_{i=1}^{m}(a_i^T x)^2(a_i^T z_\perp)(a_i^T w)\Big|.
\end{aligned}$$

First Term. Note that we have a product of independent subexponential random variables, and so if we condition on the bounds

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}(a_i^T x)^2 &< 3 \\
\max_{1\leq i\leq m}(a_i^T x)^2 &< 9\log m
\end{aligned}$$

both of which happen with probability at least $1-1/m^2$, we find via Bernstein that

$$\mathbb{P}[\big|\frac{1}{m}\sum_{i=1}^{m}(a_i^T x)^2\big((a_i^T z_\perp)^2 - \|z_\perp\|_2^2\big)\big| > \delta/6] \leq 2\exp\big(-c\min\{\frac{m\delta^2}{108}, \frac{m\delta}{54\log m}\}\big)$$

which we can make smaller than $e^{-\alpha r n}$ so long as $m \geq C\delta^{-2}\alpha n r \log(n)$ and we conclude via an ϵ-net argument that

$$\mathbb{P}[\Big\|\frac{1}{m}\sum_{i=1}^{m} a_i a_i^T \otimes (a_i a_i^T P_\perp) - P_\perp\Big\|_{\big|_{X\otimes\mathbb{R}^n}} > \delta/3] \leq e^{-\beta r n} + 2/m^2 \tag{2.32}$$

where $P_\perp$ is the orthogonal projection onto the complement of X in $\mathbb{R}^n$.

Third Term. We recognize $(a_i^T z_\perp)$ as a sub-gaussian random variable, and thus if we condition on

$$\frac{1}{m}\sum_{i=1}^{m}(a_i^T x)^4 (a_i^T w)^2 \leq 16 \tag{2.33}$$

we find via Hoeffding that

$$\mathbb{P}[\big|\frac{2}{m}\sum_{i=1}^{m}(a_i^T x)^2 (a_i^T z_\perp)(a_i^T w)\big| > \delta/3] \leq \exp\big(1 - \frac{c\delta^2 m}{16}\big).$$

We can make this bound smaller than $e^{-\alpha r n}$ so long as $m \geq C\delta^{-2}\alpha r n$. As (2.33) happens with probability greater than $1 - 1/m^2$ we find

$$\mathbb{P}[\Big\|\frac{2}{m}\sum_{i=1}^{m} a_i a_i^T \otimes (P a_i a_i^T P_\perp)\Big\|_{\big|_{X\otimes\mathbb{R}^n}} > \delta/3] \leq e^{-\beta r n} + 1/m^2. \tag{2.34}$$

Second Term. For this term we further decompose w into its x-component and its $x_\perp$-component. We can apply the same analysis for the

first and third terms to the $x_\perp$ terms, and have only to deal with

$$\mathbb{P}[\big|\frac{(x^Tw)^2}{m}\sum_{i=1}^m (a_i^Tx)^4 - 3(x^Tw)^2\big| > \delta/3].$$

If we condition on

$$\max_{1\le i\le m}(a_i^Tx)^4 \le 81(\log m)^2$$

and note that

$$\big|3 - \mathbb{E}\big[(a_i^Tx)^4\big|(a_i^Tx)^4 \le 81(\log m)^2\big]\big| \le 1/m^2$$

we find via Hoeffding that

$$\mathbb{P}[\big|\frac{(x^Tw)^2}{m}\sum_{i=1}^m (a_i^Tx)^4 - \mathbb{E}[...]\big| > \delta/3] \le \exp\big(\frac{-\delta^2m^2}{Cm + 81(\log m)^2\delta m}\big) \quad (2.35)$$

which we can make smaller than $e^{-\alpha r}$ so long as $m \ge C\alpha\delta^{-2}r(\log r)^2$. Moreover, note that we can also make (2.35) smaller than $1/m^2$ for $m \ge C\delta^{-2}$. We conclude that

$$\mathbb{P}[\big\|\frac{1}{m}\sum_{i=1}^m a_ia_i^T\otimes a_ia_i^T - \mathbb{E}[a_ia_i^T\otimes a_ia_i^T]\big\|_{X\otimes X} > \delta/3] \le 3\min\{e^{-\beta r}, 1/m^2\} + 3/m^2. \quad (2.36)$$

Combining (2.32), (2.36), and (2.34) yields the stated result. □

Proof of Lemma 2.1.6. Note that $y_i = \sum_{k=1}^r (a_i^Tx_k)^2$, and so by Theorem 2.1.9 we find that with probability greater than $1 - 2e^{-\beta rn} - 6/m^2$

$$\left\|\frac{1}{m}\sum_{i=1}^m (a_i^Tx_k)^2a_ia_i^T - \|x_k\|_2^2Id - 2x_kx_k^T\right\|_{op} < \delta\|x_k\|_2^2$$

and so

$$\left\| \frac{1}{m} \sum_{i=1}^{m} y_i a_i a_i^T - \|X\|_F^2 Id - 2XX^T \right\|_{op} \leq \sum_{k=1}^{r} \left\| \frac{1}{m} \sum_{i=1}^{m} (a_i^T x_k)^2 a_i a_i^T - \|x_k\|_2^2 Id - 2x_k x_k^T \right\|_{op}$$

$$\leq \delta \|X\|_F^2.$$

□

Proof of Lemma 2.1.3. Note that

$$\nabla^2 f(X) = \frac{2}{m} \sum_{i=1}^{m} \left[X^T a_i a_i^T X \right] \otimes a_i a_i^T$$

and so we can use Theorem 2.1.9 to write

$$\left\| X^T \otimes Id \left(\frac{1}{m} \sum_{i=1}^{m} a_i a_i^T \otimes a_i a_i^T - \mathbb{E}[...] \right) X \otimes Id \right\|_{op} < \delta \|X\|_{op}^2.$$

□

2.2 Rank One

In the rank one setting where $x \in \mathbb{R}^n$, we can say more; suppose we receive noiseless samples of the form $y_i := (a_i^T x)^2$ for i.i.d. sub-gaussian vectors $\{a_i\}_{i=1}^m$, which we assume satisfy:

$$\mathbb{E}[a_i] = 0$$

$$\mathbb{E}[a_i a_i^T] = \Sigma \tag{2.37}$$

where Σ is the covariance matrix, which we assume is invertible. Consider the eigenvalue decomposition of the covariance matrix, $\Sigma = \sum_{k=1}^{n} v_k v_k^T$. An important quantity in our analysis will be

$$\tau(x) := \max_{1 \le k \le n} (v_k x)^2 \|\Sigma^{1/2} x\|_2^{-2}, \tag{2.38}$$

a coherence parameter for $\Sigma^{-1/2} x$, and

$$\mu_4 = \mathbb{E}[(v_k^T a_i)^4],$$

a 4th moment parameter. Note that we are assuming the transformed coordinates all have equal fourth moment.

With this setup, we then consider minimization of the random function

$$f(u) := \frac{1}{4m} \sum_{i=1}^{m} \left(y_i - (a_i^T u)^2 \right)^2. \tag{2.39}$$

Lemma 2.1.2 shows that

$$\nabla^2 f(u) = \frac{1}{m} \sum_{i=1}^{m} \left(3(a_i^T u)^2 - y_i \right) a_i a_i^T \tag{2.40}$$

which is a random symmetric matrix function of u. Consequently we have

$$\begin{aligned} \mathbb{E}\left[\nabla^2 f(|u^T x| x) \right] &= \left(3|u^T x|^2 - 1 \right) \mathbb{E}\left[\nabla^2 f(x) \right] \\ &= \left(3|u^T x|^2 - 1 \right) \frac{2}{m} \sum_{i=1}^{m} \mathbb{E}\left[\left((a_i^T x)^2 \right) a_i a_i^T \right] \\ &\succeq 0. \end{aligned}$$

2.2.1 Convexity

2.2.1.1 In Expectation

We first consider convexity of the function $f(u)$ defined in (2.39) (equivalently, positive semi-definiteness of the Hessian matrix $\nabla^2 f(u)$) in the neigh-

borhood of x *in expectation* with respect to the draw of a_i, or in the limit of infinitely many samples m. We first need to establish a formula for the Hessian.

Lemma 2.2.1. *Assume that* $\{a_i\}_{i=1}^m$ *are centered random vectors with* $\mathbb{E}[aa^T] = \Sigma$. *Assume further that the transformed variables* $b_i := \Sigma^{-1/2}a_i$ *have independent coordinates and equal fourth moment parameter* $\mu_4 := \mathbb{E}[b_{ik}^4]$. *Then*

$$\begin{aligned}\mathbb{E}[\nabla^2 f(u)] = &\left(3\|\Sigma^{1/2}u\|_2^2 - \|\Sigma^{1/2}x\|_2^2\right)\Sigma \\ &+ \Sigma\left(6uu^T - 2xx^T\right)\Sigma \\ &+ (\mu_4 - 3)\sum_{k=1}^n \left(3(v_k^T u)^2 - (v_k^T x)^2\right) v_k v_k^T.\end{aligned} \tag{2.41}$$

Proof. By (2.45) we only need to compute $\mathbb{E}[(a^T u)^2 aa^T]$ for an arbitrary $u \in \mathbb{R}^n$. We will consider the slightly more general expectation $\mathbb{E}[(a^T u)(a^T w)aa^T]$ for arbitrary $u, w \in \mathbb{R}^n$. Begin by assuming $\Sigma = Id$ where Id is the $n \times n$ identity matrix. Let $i, j \in [n]$ be arbitrary coordinates. We have

$$\begin{aligned}\mathbb{E}[(a^T u)(a^T w)a_i^2] &= \mathbb{E}[a_i^2 \sum_{k=1}^n a_k^2 u_k w_k + a_i^2 \sum_{k\neq j} a_k a_j u_k w_j] \\ &= \mu_4 u_i w_i + \sum_{k\neq i} u_k w_k \\ &= (\mu_4 - 1)u_i w_i + u^T w\end{aligned}$$

and for $i \neq j$,

$$\begin{aligned}\mathbb{E}[(a^T u)(a^T w)a_i a_j] &= \mathbb{E}[a_i a_j \sum_{k=1}^n a_k^2 u_k w_k + a_i a_j \sum_{k\neq l} a_k a_l u_k w_l] \\ &= u_i w_j + w_j u_i\end{aligned}$$

so that

$$\mathbb{E}[(a^T u)(a^T w)aa^T] = (u^T w)Id + uw^T + wu^T + (\mu_4 - 3)\sum_{k=1}^{n} u_k w_k e_k e_k^T. \quad (2.42)$$

If $\Sigma \neq Id$, then observe that if we define $b := \Sigma^{-1/2}a$,

$$(a^T u)^2 aa^T = \Sigma^{1/2}(b^T \Sigma^{1/2} u)^2 bb^T \Sigma^{1/2}$$

then the inner term satisfies the assumptions needs for (2.42), and so we find

$$\begin{aligned}\mathbb{E}[(a^T u)^2 aa^T] &= \Sigma^{1/2}\left(\|\Sigma^{1/2}u\|_2^2 Id + 2\Sigma^{1/2}uu^T\Sigma^{1/2} + (\mu_4 - 3)\sum_{k=1}^{n}(\Sigma^{1/2}u)_k^2 e_k e_k^T\right)\Sigma^{1/2} \\ &= \|\Sigma^{1/2}u\|_2^2\Sigma + 2\Sigma uu^T\Sigma + (\mu_4 - 3)\sum_{k=1}^{n}(v_k^T u)^2 v_k v_k^T\end{aligned}$$

where $\Sigma^{1/2} = \begin{bmatrix} v_1 & v_2 & ... & v_n \end{bmatrix}_{n\times n}$. □

We then have the following asymptotic convexity result:

Lemma 2.2.2. *Let $x \in \mathbb{R}^n$ and consider the function $\mathbb{E}[f(u)]$ with $f(u)$ defined in (2.39). Under the same assumptions as in Lemma 2.2.1 above, $\mathbb{E}[f(u)]$ is convex in the ellipse*

$$\left\{u : \|\Sigma^{1/2}(u - x)\|_2 \leq \frac{1}{3}\left(\frac{1 + \min(\tau(x), 1/2)[\mu_4 - 3]_-}{3 + \tau(x)[\mu_4 - 3]_+}\right)\|\Sigma^{1/2}x\|_2\right\}$$

where $\tau(x)$ is the coherence of $\Sigma^{1/2}x$ as in (2.38), and we have defined $[u]_- = \min\{u, 0\}$ and $[u]_+ = \max\{u, 0\}$.

The proof of Lemma 2.2.2 relies on the fact that $\mathbb{E}[\nabla^2 f(x - tw)]$ is a convex quadratic polynomial in t. The main distinction between the gaussian general sub-gaussian case is that now x is not necessarily the leading eigenvector for the Hessian. We begin with two related eigenvalue bounds:

Lemma 2.2.3. *Let $u \in \mathbb{R}^n$ be a unit vector and consider the traceless matrix*

$$Z := uu^T - \sum_{k=1}^{n} u_k^2 e_k e_k^T.$$

Suppose that

$$0 \leq u_n^2 \leq u_{n-1}^2 \leq ... \leq u_1^2.$$

Then we have

$$\lambda_{min}(Z) \geq -\min\{u_1^2, 1/2\}$$

$$\lambda_{max}(Z) \in [0, 1 - u_n^2).$$

Proof. The bound $\lambda_{min} \geq -1/2$ follows from the universal lower bound

$$\min_{\|u\|_2=1=\|v\|_2} \left(\sum_{k=1}^{n} u_k v_k \right)^2 - \sum_{k=1}^{n} u_k^2 v_k^2 \geq -1/2.$$

Otherwise, for any unit vector $v \in \mathbb{R}^n$ we have

$$\begin{aligned} \left(\sum_{k=1}^{n} u_k v_k \right)^2 - \sum_{k=1}^{n} u_k^2 v_k^2 &\geq \left(\sum_{k=1}^{n} u_k v_k \right)^2 - u_1^2 \|v\|_2^2 \\ &\geq -u_1^2 \end{aligned}$$

Finally, note that by the Gershgorin Circle Theorem, the largest eigenvalue λ_{max} is no larger than $1 - u_n^2$ and by the tracelessness of Z it must be positive.

□

Lemma 2.2.4. *Suppose that $\{a_i\}_{i=1}^m$ are centered random vectors with independent coordinates, standard covariance $\mathbb{E}[aa^T] = Id$, and equal fourth moment parameter $\mu_4 := \mathbb{E}[a_{ik}^4]$. Then*

$$\lambda_{min}\left(\mathbb{E}\left[\nabla^2 f(x)\right]\right) \geq 2\left(1 + \min\{\tau(x), 1/2\} \min\{\mu_4 - 3, 0\}\right) \|x\|_2^2 \qquad (2.43)$$

where $\tau(x) = \max\limits_{1 \leq k \leq n} \dfrac{|x_k|^2}{\|x\|_2^2}$ is the coherence of x.

Proof of Lemma 2.2.4. Suppose that $\|x\|_2 = 1$ and let

$$g(\mu) := \lambda_{min}\left(Id + 2xx^T + (\mu - 3)\sum_{k=1}^{n} x_k^2 e_k e_k^T\right).$$

Lemma 2.2.3 above shows $g(1) \geq 1 - \min\{2\tau(x), 1\}$ and it is clear that $g(3) = 1$. By concavity of $\lambda_{min}(\cdot)$ we then find

$$\begin{aligned} g(\mu) &\geq \min\{\tau(x), 1/2\}\mu + 1 - 3\min\{\tau(x), 1/2\} \\ &= 1 + \min\{\mu - 3, 0\} \end{aligned}$$

for all $\mu \in [1, 3]$.

If $\mu_4 > 3$, then $2xx^T + (\mu_4 - 3)\sum_{k=1}^{n} x_k^2 e_k e_k^T$ is a positive semi-definite matrix and thus $g(\mu) \geq 1$. Observing that

$$2g(\mu)\|x\|_2^2 = \lambda_{min}\left(\mathbb{E}\left[\nabla^2 f(x)\right]\right)$$

produces the bound (2.43). □

Proof of Lemma 2.2.2. Begin by assuming $\Sigma = Id$ and reparametrize an arbitrary $u \in \mathbb{R}^n$ as $u = x - tw$ for $\|w\|_2 = 1$. Note that

$$\nabla^2 f(x - tw) = \frac{1}{m}\sum_{i=1}^{m}\left(2(a_i^T x)^2 - 6(a_i^T x)(a_i^T w)t + 3(a_i^T w)^2 t^2\right) a_i a_i^T \quad (2.45)$$

so using (2.42) we find

$$\begin{aligned} \mathbb{E}[\nabla^2 f(x - tw)] = 3&\left[Id + 2ww^T + (\mu_4 - 3)\sum_{k=1}^{n} w_k^2 e_k e_k^T\right] t^2 \\ &- 6\left[(x^T w)Id + xw^T + wx^T + (\mu_4 - 3)\sum_{k=1}^{n} x_k w_k e_k e_k^T\right] t \\ &+ 2\left[\|x\|_2^2 Id + 2xx^T + (\mu_4 - 3)\sum_{k=1}^{n} x_k^2 e_k e_k^T\right]. \end{aligned}$$

Observe that

$$\|x\|_2^2 Id + 2xx^T + (\mu_4 - 3)\sum_{k=1}^{n} x_k^2 e_k e_k^T \succeq (1 + \min\{\tau(x), 1/2\}[\mu_4 - 3]_-)\,\|x\|_2^2 Id \tag{2.46}$$

and that

$$(x^T w)Id + xw^T + wx^T + (\mu_4 - 3)\sum_{k=1}^{n} x_k w_k e_k e_k^T \preceq (3 + \tau(x)[\mu_4 - 3]_+)\,\|x\|_2 Id. \tag{2.47}$$

For (2.46), we used Lemma 2.2.4. Lastly, note that

$$Id + 2ww^T + (\mu_4 - 3)\sum_{k=1}^{n} w_k^2 e_k e_k^T \succeq 0.$$

Consequently we can define the polynomials

$$Q_{y,w}(t) := y^T \mathbb{E}[\nabla^2 f(x - tw)]y$$

and by convexity we can bound the smallest positive root by the intercept of the tangent line; the bounds (2.46) and (2.47) thus yield the stated conclusion for $\Sigma = Id$.

For general covariance matrices, note that we have just shown that

$$\Sigma^{-1/2}\mathbb{E}[\nabla^2 f(u)]\Sigma^{-1/2} \succeq 0$$

whenever $\Sigma^{1/2}u$ and $\Sigma^{1/2}x$ are close enough, which implies

$$\mathbb{E}[\nabla^2 f(u)] \succeq 0.$$

□

Remark 2.2.5. It is interesting to note that we have *not* assumed sub-gaussian measurement vectors up to this point. Moreover, the results above provide enough information to prove performance guarantees for stochastic gradient descent after an initialization procedure, under *no regularity* requirements. However, to ensure uniform convexity at finite sample size $m \geq Cn\log(n)$, we will need a more refined analysis based on the structure of the Hessian matrix, as presented in the next section. This will require the sub-gaussian assumption.

2.2.1.2 Finitely Many Samples

In this section we prove the main rank one convexity result:

Theorem 2.2.6 (Strong convexity)**.** *Let $x \in \mathbb{R}^n$ and $\{a_i\}_{i=1}^m$ be i.i.d. sub-gaussian satisfying $\mathbb{E}[a_i] = 0$ and $\mathbb{E}[a_i a_i^T] = \Sigma$. Define*

$$\lambda := \lambda_{min}\left(\mathbb{E}\left[\nabla^2 f(x)\right]\right).$$

If $m \geq C\|\Sigma\|_{op}^2 n(\log n)^3$, then with probability greater than $1 - 4/n^2$,

$$(u-x)^T\nabla^2 f(u)(u-x) \geq \lambda^2/(12\|\Sigma^{1/2}x\|_2^2)$$

holds uniformly for all $u \in \mathbb{R}^n$ in the ellipse around x defined by

$$\|\Sigma^{1/2}(u-x)\|_2 \leq \frac{\lambda}{30\|\Sigma^{1/2}x\|_2}.$$

Above, $C > 0$ is a constant which depends only on the sub-gaussian norm of a_i.

For a broad class of sub-gaussian measurements, Lemma 2.2.4 provides an explicit lower bound on $\lambda_{min}\left(\mathbb{E}\left[\nabla^2 f(x)\right]\right)$, thereby establishing a quantitative bound on the strong convexity parameter.

1. **Bernoulli:** For standard Bernoulli measurement vectors, where a_{ik} are i.i.d. ± 1 with equal probability, $\mu_4 = 1$ and we have a quantifiable strong convexity guarantee so long as x is incoherent, i.e., $\tau(x) < 1/2$. This is sharp in the sense that for $x = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$ the expected Hessian has a 0 eigenvalue.

2. **Gaussian:** For vectors a_i with i.i.d. standard Gaussian entries, $\mu_4 = 3$ and Lemma 2.2.4 provides the uniform lower bound

$$\nabla^2 f(u) \succeq \frac{1}{3}\|x\|_2^2 \tag{2.49}$$

   for all $\|u - x\|_2 \leq \frac{1}{15}\|x\|_2$.

3. **Sparse Gaussian:** Note that (2.49) holds anytime $\mu_4 \geq 3$ by Lemma 2.2.4. This includes sparse Gaussian vectors, whose coordinates are i.i.d. standard normal with probability p and 0 with probability $1 - p$. In this case

$$\mu_4 = 3/p.$$

We begin with a short sketch of the idea of the proof: as before, we will use the standard concentration result:

Lemma 2.2.7. *Let $x \in \mathbb{R}^n$ and $\{a_i\}_{i=1}^m$ be i.i.d. sub-gaussian, satisfying (2.37). Then there exists a constant C depending only on the sub-gaussian norm of a_i such that if $m \geq C\epsilon^{-2}\|\Sigma\|_{op}^2 n(\log n)^3$, then with probability greater than $1 - 3/n^2$ it holds that*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (a_i^T x)^2 a_i a_i^T - \mathbb{E}\left[(a^T x)^2 a a^T\right] \right\|_{op} < \epsilon \|\Sigma^{1/2} x\|_2^2.$$

This result can be proved by first truncating the norms of the measurements vectors and then applying Matrix Bernstein's Inequality (e.g., [73]). The sampling complexity can be improved, but we state Lemma 2.2.7 as a general proof-of-concept. For details see §2.2.3. For ϵ sufficiently small, this result indicates that we can control the eigenvalues of $\nabla^2 f(u)$ for u sufficiently close to x. In particular, if $\nabla^2 f(u)$ is positive definite in a region around x, then $f(u)$ is strongly convex and x is the unique minimum in this region. It is not immediately clear how to extend such control to a *quantifiable* region around x.

Assuming that $\Sigma = Id$, the same technique from §2.1.1 can be applied: first write $u = x + t\hat{w}$ for a unit vector $\|\hat{w}\|_2 = 1$ and observe that

$$\begin{aligned}(u-x)^T \nabla^2 f(u)(u-x) &= \frac{1}{m} \sum_{i=1}^{m} 3(a_i^T \hat{w})^4 t^2 + 6(a_i^T \hat{w})^3 (a_i^T x) t + 2(a_i^T x \hat{w}^T a_i)^2 \\ &= \frac{3}{m} \sum_{i=1}^{m} (A_i t + B_i)^2 - \frac{1}{m} \sum_{i=1}^{m} B_i^2\end{aligned}$$

where we have defined

$$\begin{aligned}A_i &:= (a_i^T \hat{w})^2 \\ B_i &:= \left(a_i^T x \hat{w}^T a_i\right).\end{aligned}$$

Note that

$$\frac{1}{m}\sum_{i=1}^{m} B_i^2 = \frac{1}{2}(u-x)^T\nabla^2 f(x)(u-x)$$

and consequently using Lemma 2.2.7 and Lemma 2.2.2 we can control this term. As before, applying Lemma 2.1.5 to the positive term

$$\frac{3}{m}\sum_{i=1}^{m}(A_i t + B_i)^2$$

yields the stated conclusion.

Proof of Theorem 2.2.6. Assume without loss that $\|x\|_2 = 1$ and that $\hat{w}$ is the normalized direction from u to x. Moreover begin by assuming $\Sigma = Id$. Closely following the proof of Theorem 2.1.4 we first note that

$$f''(t) = \frac{1}{m}\sum_{i=1}^{m} 3(a_i^T\hat{w})^4 t^2 + 6(a_i^T\hat{w})^3(a_i^T x)t + 2(a_i^T x\hat{w}^T a_i)^2$$

and define

$$Z_i := (A_i t + B_i)^2 \geq 0$$

where

$$A_i := \left(a_i^T\hat{w}\hat{w}^T a_i\right)$$
$$B_i := \left(a_i^T x\hat{w}^T a_i\right).$$

Note that by the computations done in Lemma 2.2.1 we have

$$\begin{aligned}
\mu(t) &:= \mathbb{E}[Z_i(t)] \\
&= \left(3 + (\mu_4 - 3)\|\hat{w}\|_4^4\right)t^2 + 2\left(3x^T\hat{w} + (\mu_4 - 3)\sum_k x_k\hat{w}_k^3\right)t + \frac{1}{2}\hat{w}^T\mathbb{E}[\nabla^2 f(x)]\hat{w} \\
&\leq \left(3 + [\mu_4 - 3]_+\right)t^2 + \left(6 + 2[\mu_4 - 3]_+\right)t + \frac{3 + 2[\mu_4 - 3]_+}{2}
\end{aligned}$$

and that the variance

$$\mathbb{E}[(Z_i(t) - \mu(t))^2] \leq C^2(t)$$

where $C(t)$ depends only on the subgaussian norm of the a_i.

Now, for a given $\epsilon \in (0, 1)$ define

$$\begin{aligned}
Y_i(t) &:= \mu(t) - Z_i(t) \\
b(t) &:= (3 + [\mu_4 - 3]_+)\, t^2 + (6 + 2[\mu_4 - 3]_+)\, t + \frac{3 + 2[\mu_4 - 3]_+}{2} \\
v^2(t) &:= C(t)^2 \\
\sigma^2(t) &:= mC(t)^2 \\
y &:= m\lambda_{min}(\mathbb{E}[\nabla^2 f(x)])/12 = m\lambda/12.
\end{aligned}$$

Applying Lemma 2.1.5 above yields

$$\mathbb{P}\left[\mu(t) - \frac{1}{m}\sum_{i=1}^{m} Z_i(t) \geq \lambda/12\right] \leq \min\left\{\exp\left(-\frac{\lambda^2 m}{144C(t)^2}\right), 25\left(1 - \Phi(\frac{\lambda\sqrt{m}}{12C(t)})\right)\right\}.$$

Using the well-known bound

$$1 - \Phi(\frac{\lambda\sqrt{m}}{12C(t)}) < \frac{12C(t)}{\lambda\sqrt{2m\pi}} \exp\left(-\frac{\lambda^2 m}{288C(t)^2}\right)$$

we find that if $m \geq 288\alpha\lambda^{-2}C(t)^2 n$ then with probability at least $1 - e^{-\alpha n}$ we

have

$$\begin{aligned} f''(t) &= \frac{3}{m}\sum_{i=1}^{m}(A_i t + B_i)^2 - \frac{1}{m}\sum_{i=1}^{m} B_i^2 \\ &\geq 3\mu(t) - \lambda/4 - \frac{1}{m}\sum_{i=1}^{m} B_i^2 \\ &\geq (9 + 3[\mu_4 - 3]_-)t^2 + (6[\mu_4 - 3]_- - 3)t + \hat{w}^T \mathbb{E}\left[\nabla^2 f(x)\right]\hat{w} \\ &\quad + \frac{1}{2}\hat{w}^T \mathbb{E}\left[\nabla^2 f(x)\right]\hat{w} - \lambda/4 - \frac{1}{2}\hat{w}^T \nabla^2 f(x)\hat{w} \\ &\geq (9 + 3[\mu_4 - 3]_-)t^2 + (6[\mu_4 - 3]_- - 3)t + \hat{w}^T \mathbb{E}\left[\nabla^2 f(x)\right]\hat{w} - \lambda/4 - \epsilon/2 \\ &\geq (9 + 3[\mu_4 - 3]_-)t^2 + (6[\mu_4 - 3]_- - 3)t + \lambda/2 \end{aligned} \tag{2.53}$$

where we used the concentration guaranteed by Lemma 2.2.7 above with $\epsilon < \lambda/8$ and the fact that

$$\frac{1}{m}\sum_{i=1}^{m} B_i^2 = \frac{1}{2}\hat{w}^T \nabla^2 f(x)\hat{w}.$$

Consequently, using a tangent line bound for the smallest positive root we find that for all

$$0 \leq t < \frac{\lambda}{6 - 12[\mu_4 - 3]_-}$$

we have

$$f''(t) \geq \lambda^2/12.$$

Moreover, an ϵ-net argument over all directions $\hat{w}$ shows that (2.53) holds for an arbitrary $\hat{w}$ with probability at least $1 - e^{-\beta n}$.

For general covariance matrices, apply the previous argument to $\Sigma^{1/2}u$ and $\Sigma^{1/2}x$ with measurements $b_i = \Sigma^{-1/2}a_i$ as usual.

□

2.2.2 Initialization

For general subgaussian measurements, we can prove good initializations so long as x is not 'too peaky' or concentrated.

Theorem 2.2.8. *Let $x \in \mathbb{R}^n$, $r \in (0, \infty)$ and define*

$$\tau(x) := \max_{1 \le k \le n} (v_k^T x)^2 \|\Sigma^{1/2} x\|_2^{-2}$$
$$\rho(x) := \min_{v_k^T x \neq 0} (v_k^T x)^2 \|\Sigma^{1/2} x\|_2^{-2}.$$

Let the vectors a_i be subgaussian satisfying (2.37), and suppose further that x satisfies the following coherence assumptions:

$$\tau(x) - \rho(x) < \frac{2}{|\mu_4 - 3|} \tag{2.55a}$$
$$\left| \frac{(\mu_4 - 3)(\tau(x) - \rho(x))}{4 + (\mu_4 - 3)(\tau(x) + \rho(x)) - 2(\mu_4 - 3)_+ \tau(x)} \right| \le \frac{r}{3\sqrt{2}}. \tag{2.55b}$$

Let z_1 be the normalized eigenvector corresponding to the largest eigenvalue of the matrix

$$M := \Sigma^{-1/2} \frac{1}{m} \sum_{i=1}^m (a_i^T x)^2 a_i a_i^T \Sigma^{-1/2}$$

and let

$$\lambda := \frac{1}{m} \sum_{i=1}^m (a_i^T x)^2$$
$$x_0 = \Sigma^{-1/2} \sqrt{\lambda} z_1.$$

If $m \ge \|\Sigma\|_{op}^2 C r^{-2} n (\log n)^3$, then the following holds with probability exceeding $1 - e^{-cn} - 3/n^2$:

$$\left\| \Sigma^{1/2} \left(x_0 - \operatorname{sgn}(x_0^T x) x \right) \right\| < r \|\Sigma^{1/2} x\|_2.$$

The constant C appearing above depends only on the subgaussian norm of a_i and on the amount of slack appearing in (2.55a).

The coherence assumptions (2.55a) and (2.55b) appear due to the fact that x is not necessarily the leading eigenvector of the expected Hessian (2.41); these assumptions guarantee that x is at least close to the leading eigenvector.

We begin with a lemma:

Lemma 2.2.9. *Let $u = (u_k)_{k=1}^n \in \mathbb{R}^n$ satisfy*

$$0 \leq u_n^2 \leq ... \leq u_1^2$$

and for some $\alpha \in [-1, \infty)$, consider the matrix

$$Z := uu^T + \alpha \sum_{k=1}^n u_k^2 e_k e_k^T.$$

Let $u_i^2 = \min_{u_k \neq 0} u_k^2$ and assume that $\alpha \leq \frac{2}{u_1^2 - u_i^2}$. Let v be the normalized eigenvector corresponding to the largest magnitude eigenvalue of Z. Let $\theta = \cos^{-1}\left(\left(\frac{u}{\|u\|}\right)^T v\right)$ be the angle between v and $\frac{u}{\|u\|}$. Then we have that

$$\sin\theta \leq \left| \frac{(\alpha/2)(u_1^2 - u_i^2)}{\|u\|_2^2 + (\alpha/2)(u_1^2 + u_i^2) - \alpha_- u_n^2 - \alpha_+ u_1^2} \right|.$$

Proof. Dividing Z by $\|u\|_2^2$ does not affect u nor the claim, so assume without loss that $\|u\|_2 = 1$. In this case, we know the largest eigenvalue of is positive as $\alpha \geq -1$. Note that we can rewrite Z as

$$\begin{aligned} Z &= \left[uu^T + \frac{\alpha}{2} \sum_{k=i}^n (u_1^2 + u_i^2) e_k e_k^T \right] + \left[\alpha \sum_{k=1}^n u_k^2 e_k e_k^T - \frac{\alpha}{2} \sum_{k=i}^n (u_1^2 + u_i^2) e_k e_k^T \right] \\ &=: D + E. \end{aligned}$$

Note that the eigenvector corresponding to the largest eigenvalue of D is u, with corresponding eigenvalue $1 + \frac{\alpha}{2}(u_1^2 + u_i^2)$. E is a diagonal matrix so we have

$$\|E\|_{op} = \left|\frac{\alpha}{2}\right| (u_1^2 - u_i^2).$$

Lastly, note that by Lemma 2.2.3 we know

$$\begin{cases} \lambda_2(Z) \in [\alpha u_{n-1}^2, \alpha u_n^2] & \text{if } \alpha < 0 \\ \lambda_2(Z) \in [\alpha u_2^2, \alpha u_1^2] & \text{if } \alpha \geq 0 \end{cases}$$

so that

$$\left|1 + \frac{\alpha}{2}(u_1^2 + u_i^2) - \lambda_2(Z)\right| \geq \left|1 + \frac{\alpha}{2}(u_1^2 + u_i^2) - \alpha_- u_n^2 - \alpha_+ u_1^2\right|$$

where here we have used our assumption on the upper bound on α. Combining all of these facts and applying the Davis-Kahan $\sin\theta$ theorem yields

$$\begin{aligned} \sin\theta &\leq \frac{\|E\|_{op}}{\left|1 + \frac{\alpha}{2}(u_1^2 + u_i^2) - \lambda_2(Z)\right|} \\ &\leq \left|\frac{\frac{\alpha}{2}(u_1^2 - u_i^2)}{1 + \frac{\alpha}{2}(u_1^2 + u_i^2) - \alpha_- u_n^2 - \alpha_+ u_1^2}\right| \end{aligned}$$

which is the stated conclusion. □

Proof of Theorem 2.2.8. Suppose for notational simplicity that x_n^2 is the minimal squared coordinate and x_1^2 the maximal, and let $\alpha := (\mu_4 - 3)/2 \geq -1$. Note that Lemma 2.2.7 guarantees that with high probability we have

$$\|M - \mathbb{E}[M]\|_{op} < \frac{2 + 2\alpha x_i^2 - 2\alpha_- x_n^2 - 2\alpha_+ x_1^2}{1 + 3\sqrt{2}/r}\|x\|_2^2 \tag{2.57}$$

so long as $m \geq Cr^{-2}n(\log n)^3$, where assumption (2.55a) guarantees the RHS is positive. Moreover, by isotropy we are also guaranteed

$$\left|\lambda - \|x\|_2^2\right| < \frac{r}{3}\|x\|_2^2. \tag{2.58}$$

Write $x = \|x\|w$ and let h be the normalized eigenvector corresponding to the largest eigenvalue of $\mathbb{E}[M]$. Observe that we have

$$\begin{aligned}\left\|\sqrt{\lambda}u - \|x\|_2 w\right\|_2 &= \left|\sqrt{\lambda} - \|x\|_2\right| + \|x\|_2\|u - w\|_2 \\ &\leq \left|\frac{\lambda - \|x\|_2^2}{\|x\|_2}\right| + \|x\|_2\left(\|u - h\|_2 + \|h - w\|_2\right)\end{aligned} \tag{2.59}$$

Note that the first term can be bounded via (2.58) to get

$$\left|\frac{\lambda - \|x\|_2^2}{\|x\|_2}\right| \leq \frac{r}{3}\|x\|_2.$$

Moreover, observe that in the notation of Lemma 2.2.9 we can write $\mathbb{E}[M] = \|x\|_2^2 Id + 2Z$ and thus

$$\begin{aligned}\|x\|_2\|h - w\|_2 &\leq \sqrt{2}\|x\|_2 \sin\angle h, w \\ &\leq \sqrt{2}\|x\|_2\left|\frac{\alpha(x_1^2 - x_i^2)}{2\|x\|_2^2 + \alpha(x_1^2 + x_i^2) - 2\alpha_- x_n^2 - 2\alpha_+ x_1^2}\right| \\ &\leq \frac{r}{3}\|x\|_2\end{aligned}$$

where we have used assumption (2.55a) for Lemma 2.2.9 and (2.55b) for the second inequality.

Now, note that we can lower bound the spectral gap $\sigma := \lambda_1 - \lambda_2$ of $\mathbb{E}[M]$ via[2]

$$2 + 2\alpha x_i^2 - 2\alpha_+ x_1^2 - 2\alpha_- x_n^2$$

which is positive under assumption (2.55a). Note that (2.57) tells us that

$$\|M - \mathbb{E}[M]\|_{op} < \frac{\sigma}{1 + 3\sqrt{2}/r}$$

[2]This follows from Lemma 2.2.9 via $\left|\lambda_1(X) - \left(1 + \frac{\alpha}{2}(x_1^2 + x_i^2)\right)\right| < \frac{\alpha}{2}(x_1^2 - x_i^2)$.

and consequently the Davis-Kahan sin θ theorem [68] tells us that

$$\begin{aligned}\|x\|_2\|u-h\|_2 &\leq \sqrt{2}\|x\|_2 \sin\angle u, h \\ &\leq \frac{\sqrt{2}r}{3\sqrt{2}}\|x\|_2 \\ &< \frac{r}{3}\|x\|_2.\end{aligned}$$

Combining all of this information into (2.59) yields

$$\left\|\sqrt{\lambda}u - \|x\|_2 w\right\|_2 < r\|x\|_2.$$

For general covariance matrices, our previous work shows us that the leading eigenvector z_1 of

$$\tilde{M} := \frac{1}{m}\sum_{i=1}^{m}(b_i^T\Sigma^{1/2}x)^2 b_i b_i^T$$

will be close to $\Sigma^{1/2}x$. Thus if we let $u := \Sigma^{-1/2}z_1$ then

$$\begin{aligned}\|\Sigma^{1/2}\sqrt{\lambda}u - \Sigma^{1/2}x\|_2 &= \|\sqrt{\lambda}z_1 - \Sigma^{1/2}x\|_2 \\ &\leq r\|\Sigma^{1/2}x\|_2.\end{aligned}$$

□

2.2.3 Concentration

Proof of Lemma 2.2.7. Begin by assuming $\Sigma = Id$ and $\|x\|_2 = 1$. Note that because of the sub-gaussian assumption we have that for $m \geq C$

$$\begin{aligned}\mathbb{P}\left[(a_i^T x)^2 \geq c\log m\right] &\leq \exp\left(1 - \hat{c}\log m\right) \\ &\leq m^{-4}\end{aligned}$$

$$\begin{aligned}\mathbb{P}\left[\|a_i\|_2^2 \geq cn\log m\right] &\leq 2\exp\left(-\hat{c}\log m\right) \\ &\leq m^{-4}\end{aligned}$$

where the constants depend on the sub-gaussian norm of a_i. Consequently

$$\mathbb{P}\left[\max_{1\leq i\leq m}(a_i^T x)^2\|a_i\|_2^2 \geq c^2 n(\log m)^2\right] \leq 2m(m^{-4}) \leq 2m^{-3}$$

and if we define the truncated random variables $\tilde{a}_i := a_i\chi_{(a_i^T x)^2\|a_i\|_2^2\leq c^2 n(\log m)^2}$ and the analogous truncated matrix

$$\tilde{M} = \frac{1}{m}\sum_{i=1}^{m}(\tilde{a}_i^T x)^2\tilde{a}_i\tilde{a}_i^T,$$

Bernstein's inequality (Theorem 4.1 in [73]) tells us that

$$\mathbb{P}\left[\left\|\tilde{M} - \mathbb{E}[\tilde{M}]\right\|_{op} \geq \delta\right] \leq n\exp\left(\frac{-\delta^2 m^2/2}{\sigma^2 + mn(\log m)^2\delta/3}\right)$$

where σ^2 is given by

$$\begin{aligned}\sigma^2 &:= \left\|\sum_{i=1}^{m}\mathbb{E}\left[(\tilde{a}_i^T x)^4\|\tilde{a}_i\|_2^2\tilde{a}_i\tilde{a}_i^T\right] - \left(\mathbb{E}\left[(\tilde{a}_i^T x)^2\tilde{a}_i\tilde{a}_i^T\right]\right)^2\right\|_{op}\\ &= m\left\|\mathbb{E}\left[(\tilde{a}^T x)^4\|\tilde{a}\|_2^2\tilde{a}\tilde{a}^T\right] - \left(\mathbb{E}\left[(\tilde{a}^T x)^2\tilde{a}\tilde{a}^T\right]\right)^2\right\|_{op}\\ &\leq Cmn\end{aligned} \tag{2.60}$$

where C is a constant which depends on the moments of a_i.

Lastly observe that if $\hat{a}_i := a_i\chi_{(a_i^T x)^2\|a_i\|_2^2\geq c^2 n(\log m)^2}$ then we can write

$$\begin{aligned}\left\|\mathbb{E}[\tilde{M}] - \mathbb{E}[M]\right\|_{op} &= \left\|\mathbb{E}\left[(\hat{a}^T x)^2\hat{a}\hat{a}^T\right]\right\|_{op}\\ &\leq \mathbb{E}\left[(\hat{a}^T x)^2\|\hat{a}\|_2^2\right]\\ &= \int_0^{c^2 n(\log m)^2}\mathbb{P}\left[(a^T x)^2\|a\|_2^2 \geq c^2 n(\log m)^2\right]\ dt\\ &\quad + \int_{c^2 n(\log m)^2}^{\infty}\mathbb{P}\left[(a^T x)^2\|a\|_2^2 \geq t\right]\ dt\\ &= c^2 n(\log m)^2\Big(\mathbb{P}\left[(a^T x)^2\|a\|_2^2 \geq c^2 n(\log m)^2\right]\\ &\quad + \int_1^{\infty}\mathbb{P}\left[(a^T x)^2\|a\|_2^2 \geq \alpha c^2 n(\log m)^2\right]\ d\alpha\Big)\\ &\leq 3/m^2\end{aligned} \tag{2.61}$$

where we used Jensen's inequality for the first line and have assumed $m \geq Cn$. Consequently we find that for $m \geq Cn$

$$\begin{aligned}\mathbb{P}\left[\|M - \mathbb{E}[M]\|_{op} \geq \epsilon\right] &\leq \mathbb{P}\left[M \neq \tilde{M}\right] + \mathbb{P}\left[\left\|\tilde{M} - \mathbb{E}[M]\right\|_{op} \geq \epsilon\right] \\ &\leq m(2m^{-4}) + n\exp\left(\frac{-\delta^2 m^2/2}{\sigma^2 + mn(\log m)^2\delta/3}\right)\end{aligned}$$

where $\delta := \epsilon - 3/m^2$ from (2.61). Now, all we have left is to show that the exponential can be made less than a power of m. Using (2.60) we find that we need m to satisfy

$$m \geq n\left(\log n + 2\log m\right) \cdot \left(C + (\log m)^2\delta\right)/\delta^2$$

for which it suffices to require $m \geq C\epsilon^{-2}n(\log n)^3$ for some constant C which only depends on the moments of a_i.

For the more general statement note that our previous work shows

$$\left\|\frac{1}{m}\sum_{i=1}^{m}(b_i^T\Sigma^{1/2}x)^2 b_i b_i^T - \mathbb{E}\left[(b^T\Sigma^{1/2}x)^2 bb^T\right]\right\|_{op} < \epsilon\|\Sigma\|_{op}^{-1}\|\Sigma^{1/2}x\|_2^2$$

whenever $m \geq C\epsilon^{-2}\|\Sigma\|_{op}^2 n(\log n)^3$. Consequently,

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i=1}^{m}(a_i^T x)^2 a_i a_i^T - \mathbb{E}\left[(a^T x)^2 aa^T\right]\right\|_{op} &= \left\|\Sigma^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}(b_i^T\Sigma^{1/2}x)^2 b_i b_i^T - \mathbb{E}\left[(b^T\Sigma^{1/2}x)^2 bb^T\right]\right)\Sigma^{1/2}\right. \\ &\leq \|\Sigma\|_{op}\left\|\frac{1}{m}\sum_{i=1}^{m}(b_i^T\Sigma^{1/2}x)^2 b_i b_i^T - \mathbb{E}\left[(b^T\Sigma^{1/2}x)^2 bb^T\right]\right\|_{op} \\ &< \epsilon\|\Sigma^{1/2}x\|_2^2\end{aligned}$$

which is the desired claim. □

Chapter 3

Quadratic Non-Negative Matrix Factorization

In this chapter, we consider the family of problems (1.9), (1.10), (1.11); recall that all of these problems can be described by the optimization problem

$$\max \sum_{i=1}^{k} \rho(S_{ii}) \tag{3.1}$$

where S is a given nonnegative, irreducible matrix, $\rho(S_{ii})$ denotes the spectral radius of the principal submatrix S_{ii} and the maximum is taken over all disjoint principal submatrices of S.

This chapter is outlined as follows: in §3.1 we begin with background and the initial motivation for considering problems of this form; we will then focus on the special case where S is the transition matrix for a finite state Markov Chain in §3.2, and the mathematical quantities have a particularly intuitive meaning. In §3.3 we move on to analyzing convergence guarantees of two algorithms for solving (3.1).

3.1 Background and Setup

Given a graph $G = (V, E)$ with non-negative edge weights $\{w_e\}_{e \in E}$, we consider the problem of "optimally" partitioning the vertex set, V into k

disjoint subsets. This *graph partitioning problem* frequently arises in the machine learning community, where the vertices represent observed data points, the weights represent some notion of similarity and the goal is to identify meaningful groups (*i.e.* "clusters") within the data. One difficulty arises in choosing a measure of optimality which is both computable and intuitive.

Many clustering algorithms operate on the *graph Laplacian* matrix Δ, which can be written as $\Delta_r := D^{-r}(D - W)$ where D is the degree matrix and W is the adjacency matrix for the graph. When $r = 0$, this is the *combinatorial graph Laplacian* or *unnormalized symmetric graph Laplacian* [25]; by the Gershgorin Circle Theorem it is easy to see that Δ_0 is a positive semidefinite matrix. When $r = 1$ this is the *asymmetric normalized graph Laplacian* or *random walk graph Laplacian*; this terminology is explained by observing that $D^{-1}W$ is row-stochastic and so Δ_1 is simply a shift by the identity of the transition matrix for a reversible Markov Chain on the graph. In either instance, the popular set of spectral clustering algorithms [77] compute the leading $k+1$ eigenvectors of Δ_r and then perform k-means on this low-dimensional embedding of the graph.

As an alternative to spectral clustering for partitioning a given graph, in [62] the following 'Dirichlet Energy' was proposed

$$\min_{V=\amalg_{i=1}^{k} V_i} \sum_{i=1}^{k} \lambda(V_i) \tag{3.2}$$

where $V = \amalg_{i=1}^{k} V_i$ is a partitioning of the vertex set and $\lambda(V_i)$ represents the smallest eigenvalue of the principal submatrix formed from the indices V_i. In

the graph setting, we refer to $\lambda(V_i)$ and the corresponding eigenvector ψ_i^D as a *Dirichlet eigenvalue* and *Dirichlet eigenvector*, respectively.

This eigenvalue partitioning problem has an analogous geometric formulation, which serves as the motivation for its introduction. Namely, given a bounded open set $\Omega \subset \mathbb{R}^2$, or more generally a compact manifold, find the partition $\Omega = \amalg_{i=1}^k \Omega_i$ which attains

$$\inf_{\Omega=\amalg_{i=1}^k \Omega_i} \quad \sum_{i=1}^{k} \lambda(\Omega_i), \tag{3.3}$$

where $\lambda(\Omega_i)$ denotes the first Dirichlet-Laplace eigenavalue of Ω_i. Existence of optimal partitions for (3.3) in the class of quasi-open sets was proved in [11]. Subsequently, several papers have investigated (3.3) and similar problems, focusing on the regularity of partitions, properties of optimal partitions, the asymptotic behavior of optimal partitions as $k \to \infty$, and computational methods [13, 8, 9, 47, 46, 48, 63, 12]. The loss of infinitesimal scale on a graph has many consequences for the Dirichlet spectrum and partitioning problem. For example, the following statement is true in the continuum but fails on a graph: any eigenvalue of the Laplace-Dirichlet operator is also the first eigenvalue for each of the nodal domains of the eigenfunction.

Seemingly unrelated, nonnegative matrix factorization is the general algebraic problem of finding a factorization of a matrix $A = \prod_i^K N_i$ where some, or all of the N_i are constrained to be nonnegative. This type of problem naturally arises in variable selection [56, 49] and clustering [80]. One approach

for clustering applications is to solve

$$\min_{V \in \mathcal{X}} \|W - VV^T\|_F^2, \qquad \text{where } \mathcal{X} := \{V \in \mathbb{R}^{n \times k} \colon V^T V = \mathrm{Id}, V_{ij} \geq 0\}. \quad (3.4)$$

Here W is a similarity matrix constructed from the data, k is the desired number of clusters, and Id is the $k \times k$ identity matrix. Perhaps surprisingly, the following proposition shows that in certain instances the objective (3.2) is equivalent to the NMF objective (3.4), where the matrix to be factorized is the symmetrized random walk matrix on the graph.

Proposition 3.1.1. *Let $\Psi^* := \left[\psi_1^D | \cdots | \psi_k^D\right]$ be the matrix where the columns are Dirichlet eigenvectors corresponding to the solution of* (3.2) *for $r = 1$. Then*

$$D^{1/2}\Psi^* = \arg\min_{U \in \mathcal{M}} \|D^{-1/2} W D^{-1/2} - UU^T\|_F^2,$$
$$\textit{where} \quad \mathcal{M} := \{U \in \mathbb{R}^{n \times k} \colon U^T U = \mathrm{Id}, U_{ij} \geq 0\},$$

and $D = \mathrm{diag}(W\mathbf{1})$ is the degree matrix and W is the similarity weight matrix.

Proof. Let V be a collection of Dirichlet eigenvectors corresponding to some partition. Then, by definition, we have $\Delta V = V \mathrm{diag}(\vec{\lambda})$ where $\mathrm{diag}(\vec{\lambda})$ is a $k \times k$ diagonal matrix, with the Dirichlet eigenvalues along the diagonal. Moreover, V satisfies $V_{ij} \geq 0$ and $V^T D V = \mathrm{Id}$. Thus the partitioning problem (3.2) is equivalent to

$$\Lambda_k^* = \min_{V \in \mathbb{R}^{n \times k}} \mathrm{tr}(V^T D \Delta V)$$
$$\text{s.t. } V_{ij} \geq 0, \quad V^T D V = \mathrm{Id}.$$

Using the definition of the graph Laplacian, the objective function can be expanded to

$$\mathrm{tr}(V^T DV) - \mathrm{tr}(V^T WV) = \mathrm{tr}(V^T DV) - \mathrm{tr}(V^T D^{1/2} D^{-1/2} W D^{-1/2} D^{1/2} V).$$

Thus, (3.5) is equivalent to

$$\begin{aligned} \min_{V \in \mathbb{R}^{n \times k}} \ & \|D^{-1/2} W D^{-1/2} - D^{1/2} V V^T D^{1/2}\|_F^2 \\ \text{s.t. } & V_{ij} \geq 0, \quad V^T DV = \mathrm{Id}. \end{aligned}$$

After the change of variables $U := D^{1/2} V$, we arrive at the stated proposition.

□

We were not the first to connect NMF with spectral-based methods; [32] describes a connection between various spectral clustering objectives and NMF. However, the algorithm proposed in [62] for solving (3.2) is new for this NMF objective; typical approaches to quadratic NMF problems are algebraic and involve finding good convex approximations.

3.2 Finite state Markov Chains

In the case of finite state Markov Chains, we can justify the objective (3.1) further. Let P be the transition matrix for a Markov Chain, with sum-normalized *columns*[1]. So that we won't have to worry about uniqueness statements, assume that the entries of P are strictly positive. Let P_Ω be an

[1]This implies P acts on the left, and the entry P_{ij} denotes the probability of state j transitioning to state i.

arbitrary principal submatrix of P with spectral radius $\rho(P_\Omega)$ and corresponding right-eigenvector π_Ω, which we assume is ℓ_1 normalized. We begin with the following simple bounds:

Lemma 3.2.1. *Let P be an irreducible transition matrix for a reversible Markov Chain and $\Omega \subset [n]$ some subset of states. Let π be the stationary distribution for P. Let*

$$\Phi_\Omega := \frac{1}{\sum_{i\in\Omega} \pi_i} \sum_{i,j\in\Omega} P_{ij}\pi_i$$

denote the conditional probability of staying in Ω after one step, given that the random walk initializes in Ω via the stationary distribution. Let

$$\Xi_\Omega := \max_{k\in\Omega} \sum_{j\in\Omega} P_{kj}$$

be the maximal conditional probability of a walker ending in Ω given that it started at a particular state in Ω. Then

$$\Phi_\Omega \leq \rho(P_\Omega) \leq \Xi_\Omega.$$

Proof. The upper bound is an immediate consequence of the Gershgorin Circle Theorem.

The lower bound follows from reversibility; in this case the spectral radius of the principal submatrix P_Ω can be characterized by the following variational formula:

$$\rho(P_\Omega) = \max_v \frac{\langle v, P_\Omega v\rangle_\pi}{\langle v, v\rangle_\pi} \tag{3.6}$$

and we note that χ_Ω, the characteristic function for the states in Ω, is an admissible vector for (3.6). Consequently,

$$\begin{aligned}\rho(P_\Omega) &\geq \frac{\langle \chi_\Omega, P_\Omega \chi_\Omega \rangle_\pi}{\langle \chi_\Omega, \chi_\Omega \rangle_\pi} \\ &= \frac{\sum_{i\in\Omega} \chi_\Omega(i)\pi(i)[P_\Omega \chi_\Omega]_i}{\sum_{i\in\Omega} \pi_i} \\ &= \frac{1}{\sum_{i\in\Omega} \pi_i} \sum_{i,j\in\Omega} P_{ij}\pi_i.\end{aligned}$$

□

In fact, we can say more; given a probability vector $x \in \mathbb{R}^n_{\geq 0}$, define the *hitting time of the boundary* $\tau_\Omega(x)$ as the *first* time a walker which initializes in Ω via x leaves Ω.

With this setup, consider the following quantity:

$$\mathbb{P}\left[\tau_\Omega(x) > k\right]^{1/k}. \tag{3.8}$$

Observe that if we assume without loss of generality that x is supported on Ω, then (3.8) is given by

$$\mathbb{P}\left[\tau_\Omega(x) > k\right]^{1/k} = \left(\sum_{i\in\Omega} [P_\Omega^k x]_i\right)^{1/k}$$

i.e., it is the probability that the walker has remained within Ω for k steps.

Note that $P_\Omega/\rho(P_\Omega)$ is a substochastic matrix with operator norm 1. Thus for any probability vector x on Ω we have the crude bound

$$\sum_{i\in\Omega} [P_\Omega^k x]_i \leq \rho(A_\Omega)^k |\Omega|^{1/2} \|x\|_2.$$

Consequently we find that

$$\mathbb{P}\left[\tau_\Omega(x) > k\right]^{1/k} \leq \rho(A_\Omega)|\Omega|^{1/2k}\|x\|_2^{1/k}.$$

This implies

$$\limsup_{k\to\infty} \mathbb{P}\left[\tau_\Omega(x) > k\right]^{1/k} \leq \rho(A_\Omega).$$

Moreover, it can be shown that the sequence $a_k = \mathbb{P}\left[\tau_\Omega(x) > k\right]^{1/k}$ is eventually non-decreasing. Consequently the full limit

$$\lim_{k\to\infty} \mathbb{P}\left[\tau_\Omega(x) > k\right]^{1/k}$$

exists and is upper bounded by $\rho(P_\Omega)$. Lastly, observe that when $x = \pi_\Omega$ we have equality. We conclude the following theorem:

Theorem 3.2.2. *Let P_Ω be a principal submatrix of an irreducible non-negative transition matrix P. If we define the stopping time*

$$\tau_\Omega(x) := \min_{k\geq 1}\{X_k \notin \Omega \big| X_0 \sim x\}$$

as the first time a random walker which initializes via x leaves Ω, then

$$\rho(P_\Omega) = \max_{x:\sum_i x_i = 1} \lim_{k\to\infty} \mathbb{P}\left[\tau_\Omega(x) > k\right]^{1/k}.$$

This theorem says that the spectral radius of a principal submatrix of a transition matrix can be interpreted as the *exponential hitting time of the boundary* for the collection of substates represented by the principal submatrix. Theorem 3.2.2 has an analog in the continuous case which can be derived from the Feynman-Kac formula for elliptic PDEs, e.g. [41].

Consequently, we find that maximizing the objective

$$\sum_{i=1}^{k} \rho(P_{\Omega_i})$$

over disjoint subsets of states has the intuitive meaning of finding the collection of subsystems which are least likely to interact.

3.3 A Rearrangement Algorithm

3.3.1 A Non-convex Relaxation

In this section, we find a relaxation of the NMF problem (3.1) and introduce an efficient algorithm for solving the relaxed problem. Our results in this section hold under an extra assumption on the nonnegative irreducible matrix S: we will assume that the matrix S is similar to a symmetric matrix $\tilde{S}$, so without loss of generality in what follows we assume symmetry. This condition is satisfied, for example, if S is the transition matrix for a *reversible* Markov Chain. Moreover, we will assume that the diagonal of S is 0. Recall that under the assumption of symmetry, we are also solving the equivalent factorization problem (1.9).

For a given function, $\phi\colon [n] \to [0,1]$ and $\alpha > 0$, consider the relaxed energy

$$\rho^{\alpha}(\phi) := \max_{\|u\|=1} u^T S u - \alpha \|u\|^2_{(1-\phi)}, \quad \text{where} \quad \|u\|^2_f := \sum_{i\in[n]} f_i u_i^2. \tag{3.9}$$

Observe that $\rho^{\alpha}(\phi)$ in (3.9) is the leading eigenvalue of the perturbed operator $S - \alpha(1-\phi)$, and the eigenfunction, u^{α}, satisfies

$$\left[S - \alpha(1-\phi)\right] u = \rho^{\alpha}(\phi) u.$$

The maximizer u^α is unique up to a scaling and can be chosen to be strictly positive, *i.e.*, for all $i \in [n]$, $u_i^\alpha > 0$.[2] Throughout, we will take u^α to be positive with $\|u^\alpha\|_2 = 1$.

If $\phi = \chi_\Omega$ is the indicator function for the set $\Omega \subset [n]$, then we intuitively think of $\rho^\alpha(\chi_\Omega)$ as an approximation to $\rho(S_\Omega)$, the leading eigenvalue for the principal submatrix formed from Ω. The following lemma shows that this approximation is exact in the limit that $\alpha \to \infty$; moreover, as α becomes large the eigenvector corresponding to $\rho^\alpha(\chi_\Omega)$ becomes strongly localized on Ω. This relaxation is directly analogous to the "fictitious domain method" arising in continuous PDEs [9].

Lemma 3.3.1. *For $\Omega \subset [n]$, $\lim_{\alpha\to\infty} \rho^\alpha(\chi_\Omega) = \rho(S_\Omega)$ and $\lim_{\alpha\to\infty} u^\alpha(\chi_\Omega) = u(\Omega)$, where $u(\Omega)$ is the leading eigenvector of the principal submatrix S_Ω.*

Proof. A simple computation shows that $\frac{d\rho}{d\alpha} = -\|u\|_{\Omega^c}^2 < 0$, where u is the corresponding normalized eigenvector. Moreover, it is clear that $\rho^\alpha(\chi_\Omega) \geq \rho(\Omega)$. Consequently $\lim_{\alpha\to\infty} \rho^\alpha(\chi_\Omega)$ exists and satisfies $\lim_{\alpha\to\infty} \rho^\alpha(\chi_\Omega) \geq \rho(S_\Omega)$.

For the reverse inequality, observe that if we normalize all eigenvectors, then after possibly passing to a subsequence, there exists a $\tilde{u}$ such that $u^\alpha \to \tilde{u}$, and $\|\tilde{u}\|_{\Omega^c}^2 = 0$. Thus $\tilde{u}$ is admissible for the eigenvalue problem of the principal submatrix S_Ω, giving us that $\rho(S_\Omega) \geq \lim_{\alpha\to\infty} \rho^\alpha(\chi_\Omega)$.

[2] These facts can be obtained by observing that the perturbed matrix is a Metzler matrix, and applying the Perron Frobenius Theorem.

Since the minimizer of the Dirichlet problem is unique, $\tilde{u} = u(\Omega)$. Thus, the previous argument shows that the only limit point of $\{u^\alpha\}_\alpha$ is $u(\Omega)$ and so $\lim_{\alpha\to\infty} u^\alpha = u(\Omega)$. □

Define the admissible class

$$\mathcal{A}_k = \{\{\phi_i\}_{i=1}^k \colon \phi_i \colon [n] \to [0,1] \text{ and } \sum_{i=1}^k \phi_i = 1\}.$$

Observe that the set of indicator functions for any k-partition of the indices is a member of $\mathcal{A}_k$. For $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$ and $\alpha > 0$, we define the *relaxed energy*, $\Lambda_k^\alpha(\{\phi_\ell\}_{\ell=1}^k) = \sum_{i=1}^k \rho^\alpha(\phi_i)$, where ρ^α is defined in (3.9). Thus, a relaxed version of the eigenvalue optimization problem (1.10) can be formulated

$$\Lambda_k^{\alpha,*} := \max_{\{\phi_i\}_{i=1}^k \in \mathcal{A}_k} \sum_{i=1}^k \rho^\alpha(\phi_i). \tag{3.10}$$

It is a consequence of Lemma 3.3.1 that for any $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$, $\Lambda_k^\alpha(\{\phi_i\}_{i=1}^k)$ is monotonically decreasing in α and for any partition $[n] = \amalg_{i=1}^k V_i$, $\lim_{\alpha\to\infty} \Lambda_k^\alpha(\{\chi_{V_i}\}_{i=1}^k) = \Lambda_k(\amalg_{i=1}^k V_i)$. However, in practice, we desire a solution to (3.10) for finite $\alpha > 0$. We observe that Λ_k^α is bounded above by the spectral radius of S, and is being maximized over the compact set $\mathcal{A}_k$. Thus a maximizer always exists. Supposing momentarily that we are able to find it, it is not yet clear how to interpret the collection $\{\phi_i^*\}$, which attains the maximum, as an actual partition; i.e., some coordinates of ϕ_i^* might be fractional implying fractional assignments. The following theorem, which is analogous to a continuous version in [9, Thm. 2.3], tells us how this is accomplished.

Theorem 3.3.2. *Let $k \in \mathbb{Z}^+$ and $\alpha > 0$ be fixed. Every (local) maximizer of Λ_k^α over $\mathcal{A}_k$ is a collection of indicator functions.*

To prove Theorem 3.3.2, we first prove the following lemma.

Lemma 3.3.3. *For $\alpha > 0$ fixed, $\rho^\alpha(\phi)$ is a convex function of ϕ.*

Proof. Let $t \in (0,1)$ and $\phi_i \colon [n] \to \mathbb{R}$ for $i = 1, 2$. Using (3.9) and the fact that the maximum is achieved for some normalized u, we compute

$$\begin{aligned}\rho^\alpha(t\phi_1 + (1-t)\phi_2) &= u^T S u - \alpha \|u\|^2_{(1-t\phi_1-(1-t)\phi_2)} \\ &= u^T S u - t\alpha\|u\|^2_{(1-\phi_1)} - (1-t)\alpha\|u\|^2_{(1-\phi_2)} \\ &\leq t\rho^\alpha(\phi_1) + (1-t)\rho^\alpha(\phi_2).\end{aligned}$$

□

Proof of Theorem 3.3.2. Our proof closely follows the proof of [9, Thm. 2.3]. The set $\mathcal{A}_k$ is the probability simplex in $\mathbb{R}^k$, and its extreme points are clearly given by the indicator functions. As Lemma 3.3.3 shows, Λ^α is a convex function on $\mathcal{A}_k$, so has a maximum and at least one maximizer is an extreme point of $\mathcal{A}_k$. We now show that, in fact, *every* maximizer is an extreme point.

To this end suppose that there exists some $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$ that achieves the maximum and is not an extreme point. Since $\sum_{i=1}^k \phi_i = 1$ there exist at least two ϕ's which are not $\{0,1\}$-valued at an index $v \in [n]$. After re-indexing, suppose these are given by ϕ_1 and ϕ_2. Thus, there exists $\epsilon > 0$ such

that $\epsilon < \phi_i(v) < 1 - \epsilon \quad i = 1, 2$. By convexity of ρ^α, we have

$$\rho^\alpha(\phi_1) \leq \frac{1}{2}\rho^\alpha(\phi_1 + \epsilon 1_v) + \frac{1}{2}\rho^\alpha(\phi_1 - \epsilon 1_v)$$
$$\rho^\alpha(\phi_2) \leq \frac{1}{2}\rho^\alpha(\phi_2 + \epsilon 1_v) + \frac{1}{2}\rho^\alpha(\phi_2 - \epsilon 1_v)$$

Adding these, and recognizing the right-hand side as an average, we must have

$$\rho^\alpha(\phi_1)+\rho^\alpha(\phi_2) \leq \max\{\rho^\alpha(\phi_1+\epsilon 1_v)+\rho^\alpha(\phi_2-\epsilon 1_v), \rho^\alpha(\phi_1-\epsilon 1_v)+\rho^\alpha(\phi_2+\epsilon 1_v)\}$$

But both terms in the maximum are feasible perturbations, thus by optimality of $\{\phi_i\}$, we must have equality:

$$\rho^\alpha(\phi_1) + \rho^\alpha(\phi_2) = \rho^\alpha(\phi_1 + \epsilon 1_v) + \rho^\alpha(\phi_2 - \epsilon 1_v) = \rho^\alpha(\phi_1 - \epsilon 1_v) + \rho^\alpha(\phi_2 + \epsilon 1_v)$$

But this implies equality in (3.11) as well:

$$\rho^\alpha(\phi_1) = \tfrac{1}{2}\rho^\alpha(\phi_1 + \epsilon 1_v) + \tfrac{1}{2}\rho^\alpha(\phi_1 - \epsilon 1_v)$$

$$\rho^\alpha(\phi_2) = \tfrac{1}{2}\rho^\alpha(\phi_2 + \epsilon 1_v) + \tfrac{1}{2}\rho^\alpha(\phi_2 - \epsilon 1_v)$$

From the proof of Lemma 3.3.3, we conclude that the eigenvector u corresponding to ϕ_1 is also an eigenvector for $\phi_1 + \epsilon 1_v$ and $\phi_1 - \epsilon 1_v$. We can subtract the following equations

$$\begin{cases} Su - \alpha(1 - \phi_1 - \epsilon 1_v)u = \rho^\alpha(\phi_1 + \epsilon 1_v)u \\ Su - \alpha(1 - \phi_1 + \epsilon 1_v)u = \rho^\alpha(\phi_1 - \epsilon 1_v)u \end{cases}$$

and using $u > 0$, simplify to yield

$$\phi_1 + \epsilon 1_v - (\phi_1 - \epsilon 1_v) = 2\epsilon 1_v \equiv C > 0$$

Algorithm 2 A rearrangement algorithm for (3.10).

Input: An initial $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$.

while not converged, **do**
  For $i = 1, \ldots, k$, compute the (positive and normalized) eigenvector u_i corresponding to $\rho^\alpha(\phi_i)$ in (3.9).
  Assign each node $v \in V$ the label $i = \arg\max_j \ u_j(v)$.
  Let $\{\phi_i\}_{i=1}^k$ be the indicator functions for the labels.
end while

for some constant C, which is clearly a contradiction. □

For fixed $\alpha > 0$, we now consider the problem of solving the relaxed partitioning problem (3.10). Since $\Lambda_k^\alpha \colon \mathcal{A}_k \to \mathbb{R}$ is Fréchet differentiable, we could apply a gradient ascent algorithm analogous to the continuous method proposed in [9]. Instead, we propose a *rearrangement algorithm* (Algorithm 2). In Lemma 3.3.4, we prove that Algorithm 2 strictly increases Λ_k^α at each iteration. This result is then strengthened in Theorem 3.3.5, to show that not only do the iterates increase the objective function, but the iterates terminate in a finite number of steps to a local maximum.

Lemma 3.3.4. *Assume $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$ is not fixed by the rearrangement algorithm (Algorithm 2). Then one iteration of the rearrangement algorithm results in a strict increase in Λ_k^α.*

Proof. Suppose $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$ is not fixed by the rearrangement algorithm and let $\{\phi_i^+\}_{i=1}^k \in \mathcal{A}_k$ be the next iterate. Let u_i denote the first (normalized,

positive) eigenvector of the operator $S - \alpha(1 - \phi_i)$. We compute

$$\Lambda^\alpha(\{\phi_i\}_{i=1}^k) = \sum \rho^\alpha(\phi_i) = \sum u_i^T S u_i - \alpha \|u_i\|^2_{(1-\phi_i)} \tag{3.12a}$$

$$\leq \sum u_i^T S u_i - \alpha \|u_i\|^2_{(1-\phi_i^+)} \tag{3.12b}$$

$$\leq \sum \rho^\alpha(\phi_i^+).$$

The inequality in (3.12a) follows from the construction of the algorithm. The inequality in (3.12b) follows from (3.9). Moreover, equality in (3.12b) holds if and only if u_i is also an eigenvector for the updated operator $S - \alpha(1 - \phi_i^+)$ for all $i = 1, \ldots, k$. From the proof of Theorem 3.3.2, we find that $\phi_i - \phi_i^+$ is a constant function for all $i = 1, \ldots, k$. This contradicts the assumption that $\{\phi_i\}_{i=1}^k$ is not fixed by the rearrangement algorithm. □

Theorem 3.3.5. *Let $\alpha > 0$. For any initialization, the rearrangement algorithm 2 terminates in a finite number of steps at a local maximum of Λ_k^α, as defined in* (3.10).

Proof. It follows from Lemma 3.3.4 and the finiteness of the space of partitions that for any initialization, the rearrangement algorithm 2 converges to a fixed point in a finite number of iterations. Thus, it suffices to show that every fixed point of the algorithm is locally optimal. Let $\{\phi_i\}_{i=1}^k$ be a fixed point of the rearrangement algorithm and let u_i denote the first (normalized, positive) eigenvector of the operator $S - \alpha(1 - \phi_i)$. The Fréchet derivative of $\Lambda_k^\alpha \colon \mathcal{A}^k \to \mathbb{R}$ in the direction $\{\delta\phi_i\}_{i=1}^k$ is written

$$\left\langle \frac{\delta \Lambda_k^\alpha}{\delta\{\phi\}}, \{\delta\phi_i\} \right\rangle = \alpha \sum_i \langle u_i^2, \delta\phi_i \rangle.$$

For $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$, any admissible perturbation can be written

$$\tilde{\phi}_i = \phi_i + \sum_{v\in[n]} t_{i,v}\chi_{\{v\}}, \qquad i = 1,\dots,k$$

for constants $t_{i,v}$, such that for every $v \in [n]$, $\sum_{i=1}^k t_{i,v} = 0$ and

$$t_{i,v}\begin{cases} \geq 0 & \text{if } \phi_i(v) = 0 \\ \leq 0 & \text{if } \phi_i(v) = 1. \end{cases}$$

Using (3.13), we compute

$$\begin{aligned}\left\langle \frac{\delta\Lambda_k^\alpha}{\delta\{\phi\}}, \{\delta\phi_i\} \right\rangle &= \alpha \sum_v \sum_i t_{i,v} u_i^2(v) \\ &\leq \alpha \sum_v \sum_i t_{i,v} u_{i^*(v)}^2(v) \\ &= 0\end{aligned}$$

where $i^*(v) = \arg\max_i u_i(v)$. This proves local optimality. □

Remark 3.3.6. We refer to algorithm 2 as a rearrangement algorithm since at each iteration, the vertex functions $\{\phi_i\}$ are rearranged to increase (3.10). These types of methods were introduced by Schwarz and Steiner and have wide applications in variational problems [53]. For example, Steiner rearrangement can be used to prove the isoperimetric inequality that the ball is the minimal perimeter domain amongst all regions of equal measure. More recently, rearrangement algorithms have been used in eigenvalue optimization problems including Krein's problem: Given an open, bounded connected domain $\Omega \subset \mathbb{R}^2$ and a prescribed amount of two materials of different density, find the distribution which minimizes the smallest frequency of the clamped drum [27, 22, 50, 51].

Algorithm 2 also shares many attributes with the Merriman, Bence, and Osher (MBO) algorithm for approximating the motion by mean curvature [58, 59, 57, 43, 36, 74].

Remark 3.3.7 (A semi-supervised extension)**.** In many clustering applications, a small percentage of the data labels are known and thus it is desirable for a clustering algorithm to have a *semi-supervised extension* that allows for the incorporation of such information. The rearrangement algorithm 2 has a natural semi-supervised variant. The label membership of a subset of the points can be fixed in the algorithm and the reader may check that all proofs of convergence remain valid. Moreover, fixing these points will force the eigenvectors to 'spread out' accordingly. We apply this variant to the MNIST handwritten digit data in Section 4.2.2.

3.3.2 A Direct Method

Let us take a closer look at the Rearrangement Algorithm 2. Assume we are bipartioning ($k = 2$), and decompose the eigenvector $u_1 = \begin{bmatrix} u_{11}^T & u_{12}^T \end{bmatrix}^T$ where u_{11} is supported on $\text{supp}(\phi_1)$ and u_{12} is supported on $\text{supp}(\phi_2)$. There is a corresponding decomposition of the matrix S given by

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

By expanding out the relationship $Su_1 - \alpha \text{diag}(\phi_2)u_1 = \rho(\phi_1)u_1$ we find

$$u_{12} = \left[\rho(\phi_1)Id + \alpha Id - S_{11}\right]^{-1} S_{21}u_{11}.$$

As $\rho(\phi_1) + \alpha > \rho(S_{11})$ we can further expand the inverse as a Neumann series

$$[\rho(\phi_1)Id + \alpha Id - S_{11}]^{-1} = (\rho(\phi_1) + \alpha)^{-1}Id + \sum_{i=1}^{\infty} S_{11}^i(\rho(\phi_1) + \alpha)^{-i-1}$$

to find that for large α

$$u_{12} \approx (\rho(\phi_1) + \alpha)^{-1}S_{21}u_{11}.$$

By Lemma 3.3.1 above we see that in the limit $\alpha \to \infty$ we have

$$\alpha u_{12} \to S_{21}u_1^* \tag{3.13}$$

where u_1^* is the true eigenvector of the principal submatrix S_{11}.

Consequently we see that we can attempt to remove the degeneracies encountered when $\alpha \gg 0$ from Algorithm 2 by rescaling via (3.13). Moreover, note that

$$S_{21}u_1^* = [Su_1^*]_2.$$

This motivates Algorithm 3 below for directly solving the problem

$$\max \sum_{i=1}^{k} \rho(S_{\Omega_i}). \tag{3.14}$$

At each step, we compute the normalized eigenvector for each principal submatrix, and then 'broadcast' this information by applying the full matrix Su_i to each eigenvector and comparing the resulting values on each index. Note that this new 'direct' method does not require a tuning parameter.

Algorithm 3 A rearrangement algorithm for (3.14).

Input: An initial partition $\{\Omega_i\}_{i=1}^k$.

while not converged, **do**
For $i = 1, \ldots, k$, compute the (positive and normalized) eigenvector u_i corresponding to S_{ii}.
Assign each node $v \in V$ the label $i = \arg\max_j [Su_j]_v$.
Let $\{\Omega_i^+\}_{i=1}^k$ be the new partition formed from these labels.
end while

3.3.2.1 Convergence

Consider the following quantity:

$$\sum_{i=1}^{k} \sum_{v \in \Omega_i} (\mathbf{s}_v^T u_i)^2 \tag{3.15}$$

where $\mathbf{s}_v$ is the vth row of S. For fixed Ω_i, the u_i that maximizes (3.15) is given by the right singular vector of S, with corresponding value

$$\sum_{i=1}^{k} \sum_{v \in \Omega_i} \rho(S_{\Omega_i})^2.$$

Moreover, for fixed u_i, it is clear that the assignments which maximize (3.15) are

$$\Omega_i^+ := \left\{ v : (\mathbf{s}_v^T u_i)^2 = \max_j (\mathbf{s}_v^T u_j)^2 \right\}.$$

When S is nonnegative, we see that it is enough to compare the values $[Su_j]_v$ instead of the magnitudes. Consequently we see that (3.15) is monotonically increasing and bounded above, hence it must converge. This does *not*, unfortunately, prove local optimality for the true objective (3.14). In fact, this algorithm is identical to an EM algorithm for subspace clustering where the subspaces are given by lines [64, 75].

In the case of a Markov transition matrix, we believe another interpretation of this direct rearrangement rule is possible via hitting times, but this is left for future work.

Chapter 4

Numerics and Experiments

4.1 Quadratic Sampling

While our theoretical guarantees are for the case of noiseless random measurements $y_i = (a_i^T x)^2$, numerical studies strongly suggest that the meta-algorithm (1) of initialize and descend is moreover stable to noise, that is, given measurements of the form $y_i = (a_i^T x)^2 + \eta_i$, the algorithm successfully returns an estimate $\widehat{x}$ up to the noise level $\frac{\|\widehat{x}-x\|_2}{\|x\|_2} \leq \|\eta\|_2$. We note that numerical simulations for initialize and descend are also promising in the complex-setting, see the reference [38]. In the real-valued setting, we present a few representative experiments here.

In the following, we consider three different measurement ensembles:

- *Bernoulli*: a_i are i.i.d. Bernoulli random vectors
- *Standard Gaussian*: a_i are i.i.d. drawn from $\mathcal{N}(0, Id_{n\times n})$.
- *Gaussian with covariance*: a_i are i.i.d drawn from $\mathcal{N}(0, \Sigma)$ with covariance matrix

$$\Sigma_{i,j} = \begin{cases} 1, & i = j \\ \frac{1}{4|i-j|}, & i \neq j \end{cases}$$

In a first experiment, we fix an n-dimensional vector x of unit norm with randomly-generated coefficients, and consider noiseless measurements $y_i = (a_i^T x)^2$. We implement the meta-algorithm 1, calling Matlab's built-in function *fminunc* to find a stationary point starting from the initialization. In the local optimization procedure, we do not provide any information to fminunc other than the function itself; by default Matlab uses a quasi-Newton method for local minimization. We run this experiment using the three different measurement ensembles above, at problem size $n = 100$ and at a number of measurements $m = 2n, 3n, \ldots, 8n$. If the solution $\widehat{x}$ recovered by the algorithm is within the tolerance $\min\{\|\widehat{x} - x\|_2, \|\widehat{x} + x\|_2\} \leq .001$, we say the algorithm has succeeded in finding the global solution. In Figure 4.1, the results of this experiment are displayed, averaged over 100 trials.

Next, we analyze numerically the stability of the meta-algorithm to additive measurement noise. For these experiments, we consider noisy measurements of the form

$$y_i = (a_i^T x)^2 + \eta_i$$

where η_i are i.i.d. mean-zero uniformly distributed, and normalized such that $\|\eta\|_2 = \mu \| \sum_i (a_i^T x)^2 \|_2$ for $\mu = .5$ (low signal to noise ratio) and $\mu = 2$ (high signal to noise ratio). We observe that the meta-algorithm is robust to such additive noise, with relative reconstruction error $\min\{\|\widehat{x} - x\|_2, \|\widehat{x} + x\|_2\}$ averaging below the signal to noise threshold. We leave a theoretical analysis of this observed noise stability to future work.

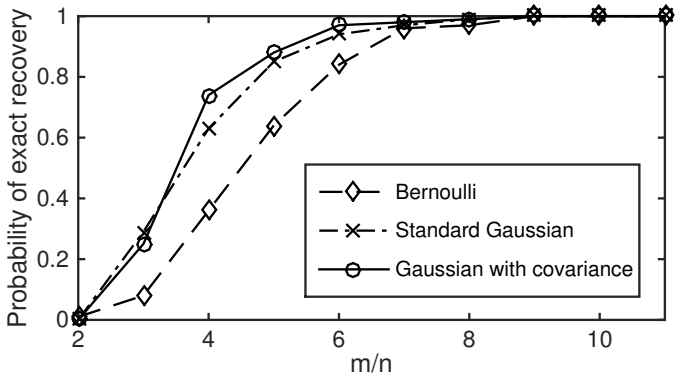

Figure 4.1: Phase transitions for exact recovery via Algorithm Initialize and Descend (1) using different random measurement ensembles.

4.2 Quadratic NMF

The results presented in this section pertain to the variant of the rearrangement algorithm for solving (3.2), as in [62].

4.2.1 Several small datasets

We obtained the similarity matrices for twelve small datasets from the website of Z. Yang [80, 79], to which we refer the reader for a complete description and source information. We apply the rearrangement algorithm to each dataset using 20 random initializations with $\alpha = k\lambda_2$. In the following table, we report the size of each dataset n, the desired number of clusters k, the purity corresponding to the lowest energy partition, a comparison value for the purity, the average number of iterations required for convergence of the rearrangement algorithm, the smallest objective function value obtained, and the objective function value of the ground truth labels. Comparison values

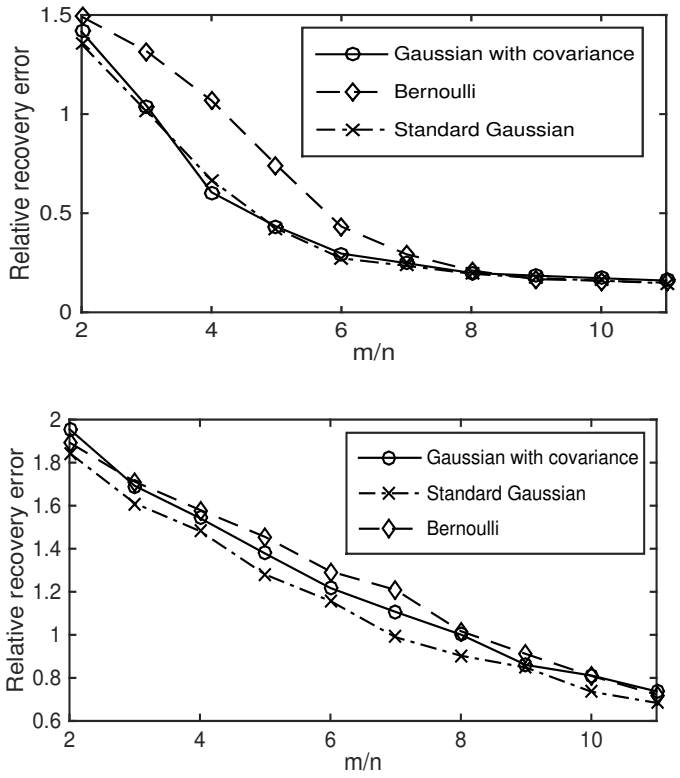

Figure 4.2: Performance of Algorithm Initialize and Descend (1) in the presence of additive noise $y_i = (a_i^T x)^2 + \eta_i$ at different signal-to-noise levels. TOP: $\|\eta\|_2 / \sum_i (a_i^T x)^2 = .5$ and BOTTOM: $\|\eta\|_2 / \sum_i (a_i^T x)^2 = 2$

for the purity are taken to be the best purity obtained by a comparison of ten different methods [80, Table 1]. It should be noted that no single method obtained the comparison purity values. We remind the reader that each iteration of the algorithm involves the computation of the ground state of k $n \times n$ standard eigenvalue problems; for these relatively small scale problems the computational costs associated with this algorithm are minimal. We observe that the found objective value is always smaller than the ground truth objective value. This demonstrates that the non-convexity of the problem is *not* preventing the algorithm from finding a good minimizer.

Dataset	n	k	purity	purity comp.	avg. iterations	found obj.	ground truth obj.
STRIKE	24	3	0.96	1.00	3.9	0.362	0.367
AMLALL	38	3	0.92	0.92	4.5	1.68	1.732
DUKE	44	2	0.52	0.70	5.2	0.549	1.019
KHAN	83	4	0.57	0.60	4.8	1.37	2.160
POLBOOKS	105	3	0.81	0.83	6.0	0.975	1.291
CANCER	198	14	0.53	0.54	5.1	13.3	16.441
SPECT	267	3	0.79	0.79	8.8	0.768	1.759
ROSETTA	300	5	0.77	0.77	7.8	5.62	12.482
ECOLI	327	5	0.80	0.83	7.0	0.568	0.656
IONOSPHERE	351	2	0.70	0.70	6.5	0.119	0.205
DIABETES	768	2	0.65	0.65	3.0	0.00552	0.013
ALPHADIGS	1404	36	0.46	0.51	12.9	37.5	56.503

4.2.2 MNIST handwritten digits

The MNIST handwritten digit dataset consists of 70,000 28×28 greyscale images of handwritten digits 0 to 9. As input we used the similarity matrix for

the MNIST dataset obtained from the website of Z. Yang [80, 79]. We symmetrize this matrix via $\tilde{W}_{ij} = \max\{W_{ij}, W_{ji}\}$, take $r = 0$, and set $\alpha = 10\lambda_2$. Moreover, we randomly sampled 3% of the data and used the semi-supervised variant of our algorithm (see §3.3.7). The remaining initialized labels were assigned randomly.

For ten different random initializations, we run the algorithm until convergence and choose the lowest energy partition. In each case, the algorithm converges in approximately 20 iterations. The purity obtained, as defined in [80], is 0.961 which is comparable to the performance of state-of-the-art clustering algorithms. We note that the partitions identified for other initial configurations had similar energy and purity values. Figure 4.3 is a graphical display of the quality of the output. On the left-hand side are the representative images for each cluster (where each eigenvector achieves its maximum), and on the right are the averaged images within each cluster. In general, the maximum value of the eigenvectors may be non-unique, but for this dataset, the maximum was unique.

Finally, we explore the real-time computational costs of the rearrangement algorithm as well as the effect of using partially labelled data. We apply the rearrangement algorithm to the MNIST dataset using 10 random initializations with $\alpha = 10\lambda_2$ and an increasing percentage of labelled data points. For each percentage of labels, we report the purity corresponding to the lowest energy partition, the average number of iterations (across initializations) required for convergence of the rearrangement algorithm, the average clock-

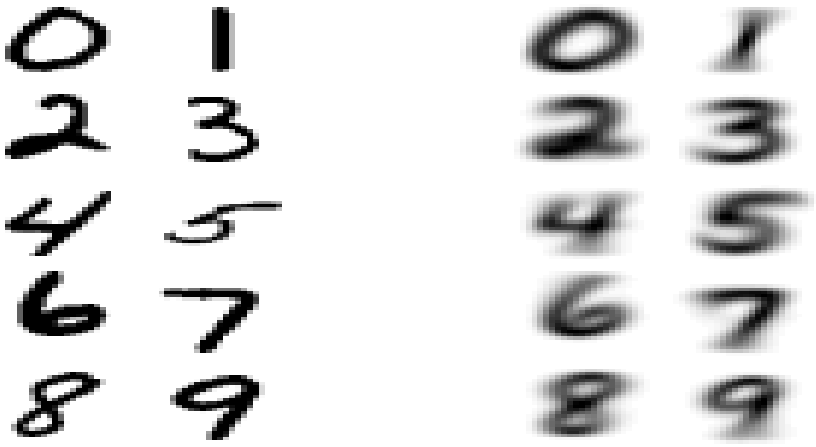

Figure 4.3: MNIST handwritten digits; each image is 28×28 pixels. **(left)** Representative images for each of the $k = 10$ clusters. **(right)** The cluster means. See §4.2.2.

time for convergence, and the smallest objective function value obtained. The objective function value for the ground truth labels is 1.6051. We observe that the average convergence times are greater than those reported in [10, Section 5], however the Dirichlet energy partitions contain additional geometric information. We also observe that for a typical random initialization, the purity is already 80-85% by the second iteration of the algorithm. All computational times reported below were obtained on a 2.67 GHz Intel Xeon desktop computer with 48GB of RAM.

MNIST dataset, $n = 70,000$**,** $k = 10$

labels	purity	avg. time (s)	avg. iterations	found obj.
1%	0.8683	115.09	29.2	1.398
2.5%	0.9619	51.15	11.4	1.563
5%	0.9702	57.27	15.3	1.565
10%	0.9711	48.22	11.5	1.567

Bibliography

[1] Boris Alexeev, Afonso S Bandeira, Matthew Fickus, and Dustin G Mixon. Phase retrieval with polarization. *SIAM Journal on Imaging Sciences*, 7(1):35–66, 2014.

[2] Pranjal Awasthi, Afonso S Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200. ACM, 2015.

[3] Radu Balan. Reconstruction of signals from magnitudes of redundant representations. *arXiv preprint arXiv:1207.1134*, 2012.

[4] Radu Balan, Bernhard G Bodmann, Peter G Casazza, and Dan Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4):488–501, 2009.

[5] Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[6] V Bentkus. An inequality for tail probabilities of martingales with differences bounded from one side. *Journal of Theoretical Probability*,

16(1):161–173, 2003.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[8] V. Bonnaillie-Noël, B. Helffer, and G. Vial. Numerical simulations for nodal domains and spectral minimal partitions. *ESAIM COCV*, 16(1):221–246, 2010.

[9] B. Bourdin, D. Bucur, and É. Oudet. Optimal partitions for eigenvalues. *SIAM Journal on Scientific Computing*, 31(6):4100–4114, 2010.

[10] X. Bresson, T. Laurent, D. Uminsky, and J. H. von Brecht. Multiclass total variation clustering. In *Advances in Neural Information Processing Systems*, pages 1421–1429, 2013.

[11] D. Bucur, G. Butazzo, and A. Henrot. Existence results for some optimal partition problems. *Adv. Math. Sci. Appl.*, 8:571–579, 1998.

[12] D. Bucur and B. Velichkov. Multiphase shape optimization problems. *arXiv:1310.2448*, 2013.

[13] L. A. Cafferelli and F. H. Lin. An optimal partition problem for eigenvalues. *J. Sci. Comp.*, 31, 2007.

[14] Emmanuel Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

[15] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Review*, 57(2):225–251, 2015.

[16] Emmanuel J Candes and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

[17] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[18] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[19] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

[20] Anwei Chai, Miguel Moscoso, and George Papanicolaou. Array imaging using intensity-only measurements. *Inverse Problems*, 27(1):015005, 2011.

[21] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM*

Journal on Optimization, 21(2):572–596, 2011.

[22] S. Chanillo, D. Grieser, M. Imai, K. Kurata, and I. Ohnishi. Symmetry breaking and other phenomena in the optimization of eigenvalues for composite membranes. *Commun. Math. Phys*, 214:315–337, 2000.

[23] Y. Chen and E. Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Arxiv preprint*, 2015.

[24] Yuxin Chen, Yuejie Chi, and Andrea Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *arXiv preprint arXiv:1310.0807*, 2013.

[25] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[26] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[27] S. J. Cox. The two phase drum with the deepest bass note. *Japan. J. Indust. Appl. Math*, 8:345–355, 1991.

[28] Gautam Dasarathy, Parikshit Shah, Badri Narayan Bhaskar, and Robert Nowak. Covariance sketching. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1026–1033. IEEE, 2012.

[29] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

[30] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Global convergence of stochastic gradient descent for some nonconvex matrix problems. *arXiv preprint arXiv:1411.1134*, 2014.

[31] Laurent Demanet and Paul Hand. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, 20(1):199–221, 2014.

[32] C. H. Q. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, pages 606–610, 2005.

[33] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[34] David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[35] Y. Eldar and S. Mendelson. Phase retrieval: stability and recovery guarantees. *Appl. and Comp. Harmon. Anal.*, 36:473–494, 2014.

[36] S. Esedoglu and F. Otto. Threshold dynamics for networks with arbitrary surface tensions. *CPAM, in press*, 2013.

[37] Ernie Esser, Michael Möller, Stanley Osher, Guillermo Sapiro, and Jack Xin. A convex model for nonnegative matrix factorization and dimensionality reduction on physical space. *Image Processing, IEEE Transactions on*, 21(7):3239–3252, 2012.

[38] M. Fickus and D. Mixon. Projection retrieval: Theory and algorithms. *Proc. SampTA*, 2015.

[39] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21:2758–2769, 1982.

[40] C. Fowlkes, S. Belongie, and J. Malik. Efficient spatiotemporal grouping using the Nyström method. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–231, 2001.

[41] Mark I Freidlin, Joseph Szücs, and Alexander D Wentzell. *Random perturbations of dynamical systems*, volume 260. Springer Science & Business Media, 2012.

[42] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[43] C. Garcia-Cardona, E. Merkurjev, A. L. Bertozzi, A. Flenner, and A. Percus. Fast multiclass segmentation using diffuse interface methods on graphs. *arXiv preprint arXiv:1302.3913*, 2013.

[44] R. W. Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35, 1972.

[45] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[46] B. Helffer. On spectral minimal partitions: A survey. *Milan J. Math.*, 78:575–590, 2010.

[47] B. Helffer and T. Hoffmann-Ostenhof. Remarks on two notions of spectral minimal partitions. *Advances in Mathematical Sciences and Applications*, 20(1):249, 2010.

[48] B. Helffer, T. Hoffmann-Ostenhof, and S. Terracini. On spectral minimal partitions: the case of the sphere. In *Around the Research of Vladimir Maz'ya III*, pages 153–178. Springer, 2010.

[49] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pages 50–57. ACM, 1999.

[50] C.-Y. Kao, Y. Lou, and E. Yanagida. Principal eigenvalue for an elliptic problem with indefinite weight on cylindrical domains. *Mathematical Biosciences and Engineering*, 5(2):315–335, 2008.

[51] C.-Y. Kao and S. Su. Efficient rearrangement algorithms for shape optimization on elliptic eigenvalue problems. *J. Sci. Comp.*, 54(2-3):492–512, 2013.

[52] Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

[53] B. Kawohl. Symmetrization-or how to prove symmetry of solutions to a PDE. *Chapman and Hall CRC Research Notes in Mathematics*, pages 214–229, 2000.

[54] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. *arXiv preprint arXiv:1410.6913*, 2014.

[55] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[56] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[57] E. Merkurjev, T. Kostic, and A. Bertozzi. An MBO scheme on graphs for segmentation and image processing. *SIAM J. Imag. Sci.*, 6(4):1903–1930, 2013.

[58] B. Merriman, J. K. Bence, and S. Osher. Diffusion generated motion by mean curvature. *UCLA CAM Report 06–32*, 1992.

[59] B. Merriman, J. K. Bence, and S. Osher. Diffusion generated motion by mean curvature. *AMS Selected Letters, Crystal Grower's Workshop*, pages 73–83, 1993.

[60] Carl D Meyer. Stochastic complementation, uncoupling markov chains, and the theory of nearly reducible systems. *SIAM review*, 31(2):240–272, 1989.

[61] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[62] Braxton Osting, Chris D White, and Édouard Oudet. Minimal dirichlet energy partitions for graphs. *SIAM Journal on Scientific Computing*, 36(4):A1635–A1651, 2014.

[63] É. Oudet. Approximation of partitions of least perimeter by γ-convergence. *Experimental Mathematics*, 20(3):260–270, 2011.

[64] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

[65] MG Raymer, M Beck, and D McAlister. Complex wave-field reconstruction using phase-space tomography. *Physical review letters*, 72(8):1137, 1994.

[66] Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pages 1214–1222, 2012.

[67] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[68] Youcef Saad. *Numerical methods for large eigenvalue problems*, volume 158. SIAM, 1992.

[69] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[70] Herbert A Simon and Albert Ando. Aggregation of variables in dynamic systems. *Econometrica: journal of the Econometric Society*, pages 111–138, 1961.

[71] Lei Tian, Justin Lee, Se Baek Oh, and George Barbastathis. Experimental compressive phase space tomography. *Optics express*, 20(8):8296–8308, 2012.

[72] Joel A Tropp. Literature survey: Nonnegative matrix factorization. *University of Texas at Asutin*, 2003.

[73] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[74] Y. van Gennip, N. Guillen, B. Osting, and A. Bertozzi. Mean curvature, threshold dynamics, and phase field theory on finite graphs. *Milan Journal of Mathematics*, 82(1):3–65, 2014.

[75] René Vidal. A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2010.

[76] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[77] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[78] Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[79] Z. Yang. personal website, accessed July 1, 2013. `http://users.ics.aalto.fi/rozyang/nmfr/NIPS2012_experiments.zip`.

[80] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja. Clustering by nonnegative matrix factorization using graph random walk. In *NIPS*, 2012.

[81] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[82] Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. *arXiv preprint arXiv:1506.07405*, 2015.

Vita

Christopher Dale White was born in Lafayette, LA in 1986. He grew up in Lake Charles, LA, attending Alfred M. Barbe High School and graduating in 2004. He then attended Louisiana State University, where he first became interested in Mathematics. After receiving a Bachelor of Science degree in Mathematics and a Bachelor of Arts degree in Economics in 2009, he then went on to enter the Ph.D. program at The University of Texas at Austin.

Permanent address: 2717 Timber Ln.
Lake Charles, LA 70605

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.